DOCUMENT RESUME

ED 119 273 CS 501 261

TITLE Status Report on Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, October 1 - December 31, 1975.
INSTITUTION Haskins Labs., New Haven, Conn.
REPORT NO SR-44-1975
PUB DATE 75
NOTE 161p.

EDRS PRICE MF-$0.83 HC-$8.69 Plus Postage
DESCRIPTORS Articulation (Speech); *Cognitive Processes; Educational Research; Higher Education; Language Development; Language Usage; *Oral Communication; *Research Methodology; *Speech; Speech Skills; *Theories
IDENTIFIERS *Status Reports

ABSTRACT

This report, covering the period from October 1 to January 31, 1975, is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. The manuscripts in this report discuss such topics as the complexity of a motor theory of speech perception, a systems approach to studying the cerebral hemispheres, the auditory and linguistic processes in speech perception, the rate at which initial phonemes are detected in spoken words and spoken nonwords, detecting nasals in continuous speech, developing a digital pattern playback for the analysis and manipulation of speech signals, the value of the voice onset time as the physical basis for separating homorganic stop categories across a variety of languages, the coarticulation tones among the Thai and using the tones as a reference system, the motor patterns that underlie articulator movements during the production of certain vowel-consonant-vowel syllables, and the findings from an ongoing study of the fricative /s/ as it is produced by normal speakers either as a single consonant or in cluster with other consonants. (RB)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SR-44 (1975)

ED119273

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 October - 31 December 1975

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

CS 501 261

CONTENTS

I. Manuscripts and Extended Reports

How Abstract Must a Motor Theory of Speech Perception Be? -- A. M. Liberman 1

A Systems Approach to the Cerebral Hemispheres -- Carol A. Fowler. . . . 17

Auditory and Linguistic Processes in Speech Perception: Inferences from Six Fusions in Dichotic Listening -- James E. Cutting. 37

Initial Phonemes Are Detected Faster in Spoken Words than in Spoken Nonwords -- Philip Rubin, M. T. Turvey, and Peter van Gelder 75

On Detecting Nasals in Continuous Speech -- Paul Mermelstein 83

A Digital Pattern Playback for the Analysis and Manipulation of Speech Signals -- P. W. Nye, L. J. Reiss, F. S. Cooper, R. M. McGuire, P. Mermelstein, and T. Montlick. 95

In (Qualified) Defense of VOT -- Leigh Lisker. 109

The Coarticulation of Tones: An Acoustic Study of Thai -- Arthur S. Abramson 119

Thai Tones as a Reference System -- Arthur S. Abramson 127

Some Electromyographic Measures of Coarticulation in VCV Utterances -- Thomas Gay 137

Durations of Articulator Movements for /s/-Stop Clusters -- Gloria J. Borden and Thomas Gay. 147

II. Publications and Reports 163

III. Appendix: DDC and ERIC numbers (SR-21/22 - SR-42/43). 165

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Assistant Chief Medical Director for Research and Development, Research Center for Prosthetics, Veterans Administration
Contract V101(134)P-342

Advanced Research Projects Agency, Information Processing Technology Office, under contract with the Office of Naval Research, Information Systems Branch
Contract N00014-67-A-0129-0002

United States Army Electronics Command, Department of Defense
Contract DAAB03-75-C-0419(L 433)

National Institute of Child Health and Human Development
Contract N01-HD-1-2420

National Institutes of Health
General Research Support Grant RR-5596

HASKINS LABORATORIES

Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper, Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
Thomas Baer*
Peter Bailey[1]
Fredericka Bell-Berti*
Gloria J. Borden*
James E. Cutting*
Ruth S. Day*
Michael F. Dorman*
Frances J. Freeman*
Jane H. Gaitenby
Thomas J. Gay*
Terry Halwes
Katherine S. Harris*
Leigh Lisker*
Ignatius G. Mattingly*
Paul Mermelstein
Seiji Niimi[2]
Lawrence J. Raphael*
Bruno H. Repp*
Philip E. Rubin*
Donald P. Shankweiler*
George N. Sholes
Michael Studdert-Kennedy*
Quentin Summerfield[1]
Michael T. Turvey*

Technical and Support Staff

Eric L. Andreasson
Dorie Baker*
Elizabeth P. Clark
Cecilia C. Dewey
Donald S. Hailey
Harriet G. Kass*
Diane Kewley-Port*
Sabina D. Koroluk
Christina R. LaColla
Roderick M. McGuire
Agnes McKeon
Terry F. Montlick
Loretta J. Reiss
William P. Scully
Richard S. Sharkany
Edward R. Wiley
David Zeichner

Students*

Mark J. Blechner
Steve Braddon
David Dechovitz
Susan Lea Donald
Donna Erickson
F. William Fischer
Carol A. Fowler
Morey J. Kitzman
Gary Kuhn
Andrea G. Levitt
Leonard Mark
Roland Mandler
Robert F. Port
Sandra Prindle
Abigail Reilly
Robert Remez
Helen Simon
Emily Tobey
Harold Tzeutschler
James M. Vigorito

---

*Part-time
[1]Visiting from The Queen's University of Belfast, Northern Ireland.
[2]Visiting from University of Tokyo, Japan.

How Abstract Must a Motor Theory of Speech Perception Be?*

A. M. Liberman+

Your kind invitation suggested that I talk about the motor theory of speech perception. Though I was reluctant to speak on that subject, I nevertheless accepted. The reason for my reluctance was that I then had nothing new to say about a motor theory, and I did not wish to rehearse the old and tired arguments. Therefore, I took the liberty of submitting an abstract that, as you may have noticed, was not exactly responsive to the invitation. Since submitting that abstract, however, my colleagues--Michael Dorman, Lawrence Raphael, Bruno Repp--and I have collected some data that are at least relevant to a motor theory. Not critical, I hasten to emphasize, only relevant; and illustrative, perhaps, of the ways one might do research on the question. These new data may also be interesting because, as we will see, they suggest that such a theory must be carefully hedged about. At all events, I decided that I could be comfortable talking once again about a motor theory if only because we do have some new data, and because these data will enable me to make explicit a restriction on the theory that has not been much discussed. Hence, I am about to take my second liberty, which is to base my talk on your invitation, after all, and not on the abstract I submitted. Even so, the abstract will be relevant, if incomplete.

Throughout this talk I will, then, be concerned with the question you wanted me to ask: Is the perception of speech linked to its production? In the first part I will say why I think that is a proper question. For that purpose I will consider what we know most generally about speech that motivates us even to wonder about a motor theory. In the second part I will describe the specific experiments by my colleagues and me that illustrate how one might go about getting an answer.

Each part will itself be organized to take into account that we should not adopt a motor theory, or any other similarly special view, until we have reason to reject an auditory theory, which is the most ordinary view. An auditory theory assumes that perceiving the phonetic message is merely an overlaid function, carried out by processes no different from those underlying the perception of music, the noises of a busy highway, or the rustle of leaves in the wind. Only if that most parsimonious theory should prove inadequate are we justified in considering some apparently less parsimonious one. It is appropriate, then, to divide the question. We should ask first: Do we have reason to suppose that

---

*This paper was delivered as a plenary address to the 8th International Congress of Phonetic Sciences, Leeds, England, 21 August 1975.

+Also University of Connecticut, Storrs, and Yale University, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

something other than auditory processes must be invoked? If the answer is yes, we may raise the second question: Can we guess what those other-than-auditory processes might be? I think that division is a particularly important one; hence I have organized what I will say so as to respect it.

Now to the general part of our discussion: What do we know about speech that motivates us to consider the questions I have just raised? Since those questions imply that the perception of speech may require its own distinctive processes, we should therefore look in speech for its most distinctive characteristic. We find that, I think, in the very peculiar nature of the relation between sound and phonetic message. The peculiar relation I speak of is a species of grammar--a speech grammar, if you will--and it is there that we should expect to find the need for special processes.

Because I will view speech as a kind of grammar, I should say a few words about grammatical codes in general so we can see where speech fits. And instead of asking about their form, which is what students of language most commonly do, I will ask rather about their function. For the moment, then, our concern is not with what these grammatical codes are, but with what they do.

To appreciate the function of grammatical codes, we need only consider the nature and shortcomings of agrammatic communication. In an agrammatic mode, which is common among animals and in man's nonlinguistic communication, the relation of message to signal is straightforward. Each message is directly linked to a signal, and every signal differs holistically from all other signals. There is no grammatical structure, only a list of all possible messages and their corresponding signals. Now if all communication were of that kind, we would not have to wonder about distinctive linguistic processes. At the one end, the signals would have to be discriminated and identified, but that is what auditory perception is all about. At the other end, the messages to which those signals are so directly connected would have to be comprehended and stored, but that is the business of processes that lie squarely in the cognitive domain. So if we knew all about auditory perception and all about cognition, we should understand the perception of agrammatic communication. No special theory would be required.

Of course, communication would be very restricted in that agrammatic world. That would be so because agrammatic communication would work as well as it needed to only if there were reasonable agreement in number between the messages people want to send and the holistically different signals they can produce and perceive. But there is the rub. The number of messages our cognitive apparatus can generate and comprehend is uncountably large, while, in contrast, our vocal tracts and ears can cope efficiently with only a small number of signals.

From a biological point of view, the existence of that mismatch is hardly to be wondered at. After all, the mismatched organs--a cognitive apparatus at the one end, a vocal tract and ear at the other--evolved separately, which is to say in connection with wholly different activities. It is tempting, then, to suppose that grammatical codes developed in evolution as a kind of interface between organs that were not made for each other. On that assumption, the function of grammar is to restructure the to-be-communicated messages so as to match the potentialities of the intellect to the limitations of the vocal tract and the ear. To the extent that the codes work well, communication becomes vastly

more efficient and various than it would otherwise be. But that gain is not to be had cheaply, since there is now a peculiar complication in the relation between signal and message, and a need for equally peculiar processes to deal with it. It is, of course, precisely those peculiar processes that would distinguish language from other psychologically interesting activities.

Given, now, that we have reason to look for special grammatical processes, we ask: What characteristics of language suggest the shape that such processes might take? Considering the problem from the standpoint of the perceiver, who is our chief concern today, I will remark the obvious: all the grammatical complications he must cope with are just those that were introduced by the speaker. It is only slightly less obvious that the same is not true for most other forms of perception. In vision, for example, the complications of shape perception are, in a very important sense, external to the perceiver. Therefore, it is reasonable, and not especially novel, to suppose about language that producing and perceiving are only different sides of the same coin. If so, the special linguistic processes that are necessary to perceive language might be expected to have something in common with those that produce it.

So much for grammatical codes in general. What about speech? One might suppose that speech is functionally different from language in that the need for a special grammatical interface has ended with the production of the phonetic message. In that case, the segments of that message would be connected to the sound in a most straightforward, agrammatical way. The perception of speech would then be no different from the perception of other sounds, and its connection to language would be only incidental. But the interfacing does not end with the phonetic message, which is only a stage in the grammatical restructuring that efficiently links meaningful message to acoustic signal. Further, and still quite drastic, changes are necessary, because the requirements of phonetic communication are not well matched to the characteristics of the vocal tract and the ear.

That mismatch has been much discussed (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman, Mattingly, and Turvey, 1973; Liberman, 1974), so I will only recall that it is possible to be quite explicit about it and thus to see clearly what the rest of the grammar--the grammar of speech--must do. For example, it is plain that if the phonetic segments were represented agrammatically in the sound, each phonetic segment by a unit sound, then we could neither speak nor listen as fast as we do. Having in mind that the phonetic message is, in fact, transmitted at rates that reach 25 or more segments per second, at least for short stretches, we see that the speeds we achieve with speech would be impossible if, as in agrammatic communication, the articulators had to change their states in step with the holistically different segments. And even if the articulators could do that, the listener's ear could not possibly resolve the unit sounds so produced: at 25 discrete acoustic segments per second, speech would become an incomprehensible buzz. Moreover, we know about auditory perception of nonspeech sounds that even at rates low enough to avoid the merging of discrete sound segments, the listener nevertheless has difficulty identifying the order in which the segments occurred. Clearly, then, there is a need for a further recoding of the information if the phonetic message is to be efficiently transmitted. And we know that such recoding does in fact occur.

It is no news to this audience that the phonetic message is not transmitted by a discrete set of articulatory gestures, one for every phonetic segment and

each in its proper turn. Rather, the segments are broken down into features; these are assigned to gestures, reorganized into units longer than a segment, and then coarticulated. That arrangement permits us to speak more rapidly than we otherwise could, since it makes for the production of phonetic segments at rates faster than we must change the states of our muscles.

But coarticulation has a perceptual function, too. Consider two of its effects on the acoustic signal and the relation of these effects to the limitations of the ear: first, information about successive segments of the message is transmitted simultaneously, so the number of acoustic segments the ear must resolve is now less than the number of phonetic segments transmitted; and, second, acoustic cues for any particular phonetic segment are different in different contexts, so the order of transmitted segments is marked, not so much by temporal order, which the ear has trouble keeping straight, but by differences in the shape of the acoustic signal. Thus we see how, by grammatical restructuring of the message, the requirements of phonetic communication are matched to the properties of the auditory system. But the fit of grammatical form to perceptual function is achieved at the cost of a complex peculiarity in the grammatical form: there is no correspondence in segmentation between message and signal, and there are odd kinds of context-conditioned variations in the acoustic cues. If listeners nevertheless cope with such peculiarities, as they do quite easily, it must be that they have access to equally peculiar processes.

Having now seen why special processes might be required for speech perception, we can ask what those processes might be. The answer would seem to be the same for speech as it was for the other grammatical codes: the complications the listener must deal with are just those that were introduced by the speaker; thus the key to the code is in the manner of its production. To adopt a motor theory is, in these terms, only to guess that the processes of perception might in some way make use of such a key.

Now I will turn, as I said I would, from general background considerations to some recent experiments bearing on the questions we have raised. First I should say that these experiments are not mine alone but are, as I indicated earlier, the results of a collaborative effort by Michael Dorman, Lawrence Raphael, Bruno Repp, and me (Dorman et al., 1975). I should add, however, my colleagues are not to be blamed for the faults of the particular interpretations I will offer today.

The experiments all have to do with a fact that is in general familiar to you namely, that the "sound of silence" is a phenomenon of speech perception and not merely a poetic image. We all hear that sound whenever we perceive stop consonants, because silence is a manner cue for those segments. Our experiments were designed to deal with that fact, and to answer four questions about it: (1) how large is the effect of the silence cue? (2) is the effect to be accounted for by the properties of the ear? (3) if not, then should we look to the vocal tract? and (4) if so, whose vocal tract?

Now to the first question: How large is the effect of silence as a cue for stops? In the first experiment the stop was in syllable-initial position after a fricative. Figure 1 gives a schematic representation of a syllable consisting of a fricative noise [ʃ], formant transitions of a type appropriate for the stop consonant [p], followed by the steady-state formants that continue the vowel [ɛ].

There was another pattern like the one in the figure, except that the stop was [k] instead of [p]. In the experiment, we varied the length of the silent interval between the noise patch and the beginning of the formant transitions. Those silent intervals ranged from 0 to 100 msec. The stimuli were real speech, not synthetic; they were presented in random order with instructions to identify each one as [ʃpɛ], [ʃkɛ], or [ʃɛ].

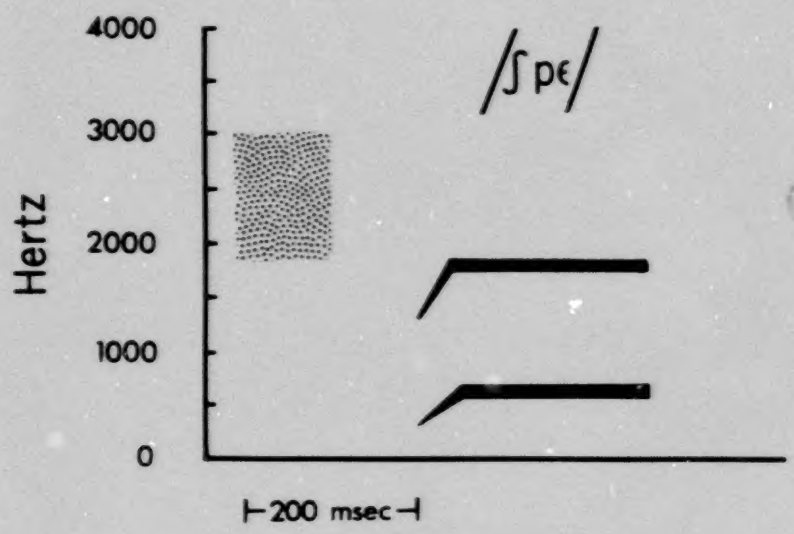


Figure 1: Schematic representation of fricative-stop-vowel syllable, illustrating the kind of pattern used to determine the effect of silence on the perception of syllable-initial stops.

Figure 2 shows the results. We see that at silent intervals of less than 20 msec our listeners reported hearing [ʃɛ], not [ʃpɛ] or [ʃkɛ]. That is, at short silent intervals the stop consonant was, for all practical purposes, not heard. The effect of the silence cue is very large indeed.

The second experiment was intended, in similar fashion, to assess the role of silence as a cue for stops, but now in syllable-final position. Figure 3 shows one of the schematic synthetic patterns that was used. There you see a two-formant disyllable [bɛb dɛ]. (The other disyllable [bɛg dɛ] is identical except that the second-formant transition at the end of the first syllable is rising instead of falling.) In these cases, the silence we are interested in is the period between the end of the first syllable and the beginning of the second. To assess its importance, we varied its duration from 0 to 120 msec and presented the stimuli for judgment as [bɛb dɛ], [bɛg dɛ], or [bɛ dɛ].

Figure 4 shows the results. We see that when the intersyllable interval is less than about 50 msec, the listeners hear [bɛ dɛ], not [bɛb dɛ] or [bɛg dɛ]. That is, with short intervals of silence, our listeners again do not hear the stop. Plainly, the effect of the silence cue is so large as to be total.

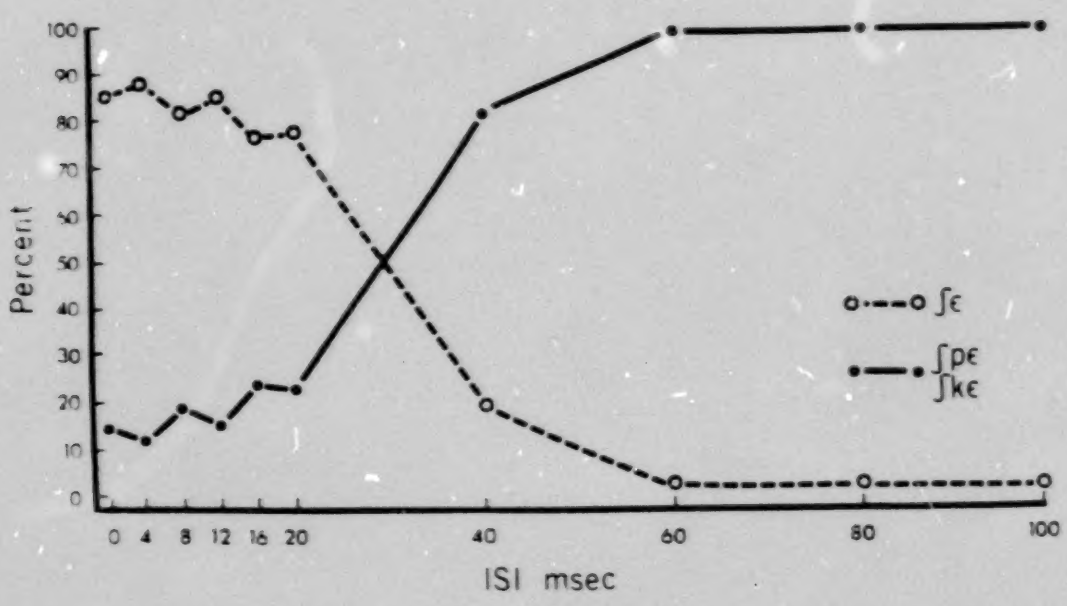


Figure 2: Percent of responses reporting the presence and absence of the stop as a function of the interval between the end of the fricative noise and the beginning of the formant transitions.

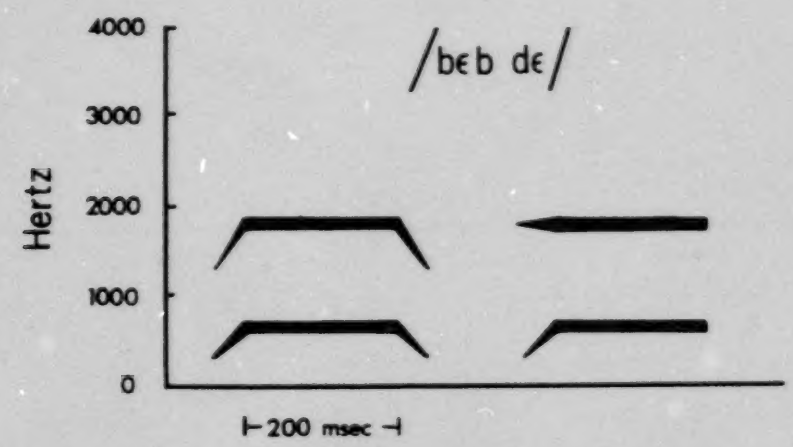


Figure 3: Schematic representation of the patterns used to determine the effect of silence on the perception of syllable-final stops.

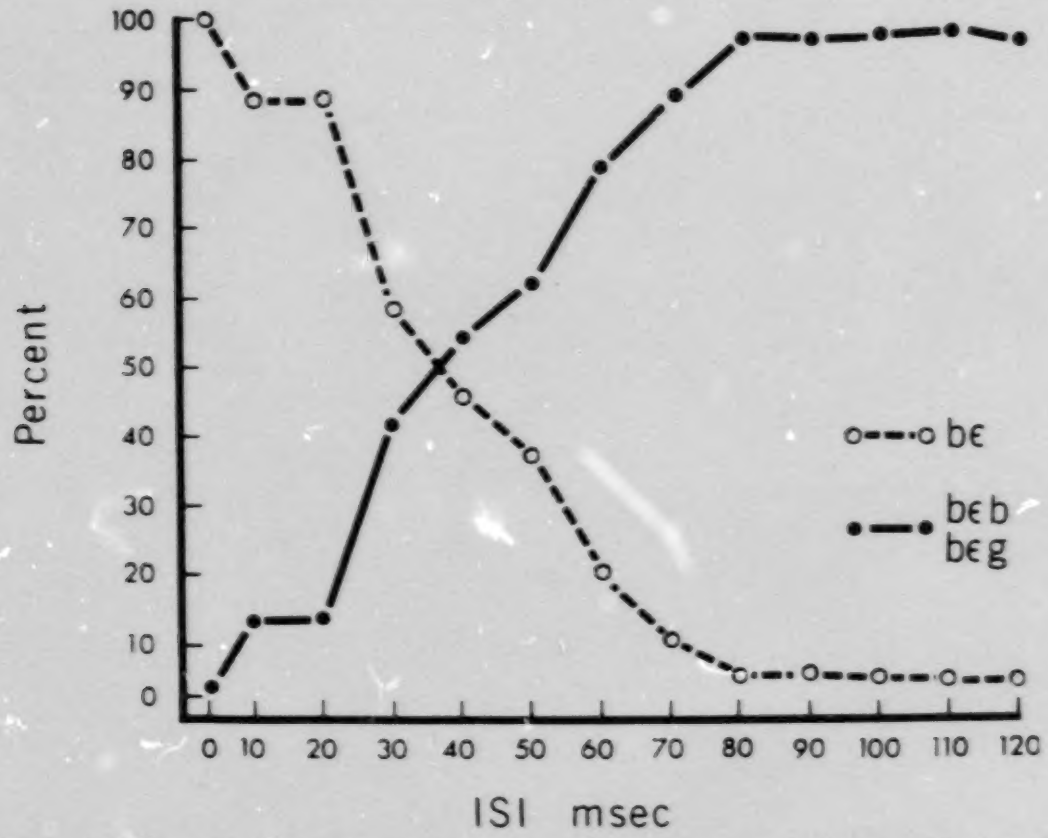


Figure 4: Percent of responses reporting the presence and absence of the stop as a function of the interval between the end of the first syllable and the beginning of the second.

But the data of those experiments only provide a background against which to ask the next question, which is more relevant to our concerns today: Can we account for the effect of the silence cue in purely auditory terms or must we look to some other-than-auditory processes? Having observed that a silent period is necessary if the stop consonant is to be heard, can we suppose that the importance of silence is owing to some general characteristic of the ear--some characteristic that has no more to do with speech than with any other sounds? Perhaps it is. In the case of [ʃpɛ] and [ʃkɛ], for example, we note that we have conformed to the paradigm for auditory forward masking. It is possible, therefore, that the noise of the fricative masks the transition cues for the stop, rendering them ineffective at an auditory level. On that account, the period of silence between fricative noise and transitions is necessary if the latter is to escape the masking effect of the former. In the case of [bɛb dɛ] and [bɛg dɛ], we have the paradigm for backward masking. Conceivably, the second syllable "backward masks" the stops at the end of the first syllable, in which case the period of silence would presumably permit the syllable-final stop to evade the masking effect.

When we examine the experimental literature, we find reasons for rejecting an auditory interpretation, especially in the syllable-initial case. Thus, a review of what is known about auditory forward masking reveals that it is typically not a large effect; there appears to be no precedent for the total masking

that we should have to assume in order to account for the disappearance of the stop consonant when the fricative noise is placed in front of it (Elliott, 1971; Leshowitz and Cudahy, 1973). More affirmatively, we find in the literature on speech perception that when the transition cues appear immediately after the fricative noise they do nevertheless have an effect--that is, they are not masked--even though they do not lead to the perception of a stop consonant. Thus, Darwin (1969) found quite incidentally that fricative-vowel syllables were more intelligible with the appropriate consonant-vowel transitions than without them. More recently, Ganong (1975) has found in an adaptation-shift experiment that the boundary between [de] and [be] was moved just as much by adaptation with [se] (that is, fricative noise followed by transitions appropriate for [se] and for [de]) as by adaptation with [de] itself. Given that the [be-de] boundary did not shift nearly so much when the transitions were removed from the adapting [se] stimulus, we may conclude that the transition cues were "getting through" effectively even though they followed close on the fricative noise.

We have aimed to get at the matter more directly. To do that in the case of the syllable-initial stops [ʃpɛ] and [ʃkɛ], we varied the silent interval after the fricative noise, exactly as we had before, but instead of the whole syllables [pɛ] or [kɛ], we presented instead only the isolated second- and third-formant transitions, which are the distinguishing cues. As you know, these isolated transitions do not sound like speech but like chirps. Fortunately for our purposes, listeners can readily learn to identify them differentially as "high" and "low." The fricative-chirp patterns were presented for judgment as "high," "low," and "no chirp."

Figure 5 shows the results. First, it reproduces in the more nearly solid line, the results of the earlier experiment. That solid curve represents the frequency with which the subjects heard [ʃɛ]--that is, the frequency with which they heard no stop. We see again that at short intervals of silence they failed to hear a stop. Now we see in the bottom line how often they failed to hear the chirp, which was never. That is, for those cases in which the listeners did not hear the stop, they nevertheless heard the essential but isolated transition cues; moreover, they heard them loud and clear.

Figure 6 shows percent correct in the perception of the stops, on the one hand, and the chirps on the other. We see, as we had before, that [pɛ] and [kɛ] are perceived correctly only when there is an interval of silence of 20 to 40 msec, but the corresponding chirps are heard correctly at all silent intervals.

A similar experiment was carried out on the stops in syllable-final position--that is, in [bɛb dɛ] and [bɛg dɛ]. The [b] and [g] transitions were isolated and placed before the syllable [dɛ] with the same intervals of silence that had been used before. Figure 7 shows the results. The more nearly solid line is from the earlier experiment and shows that the listeners heard [bɛ dɛ] at short intervals of silence. That is, at short intervals they did not hear the syllable-final stop. The dashed line at the bottom shows that the subjects never reported not hearing the chirps or, to put it affirmatively, that they always heard the chirps. Figure 8 shows how often the listeners perceived the stops and the chirps correctly. The lower, more nearly solid line reproduces the earlier result and shows that at short intervals of silence the listeners did not correctly perceive the syllable-final stops [b] and [g]. The upper, dashed line shows that at these same short intervals of silence they did perceive the chirps--that is, the transition cues--correctly, though there is perhaps a small effect of masking on the accuracy of that perception.

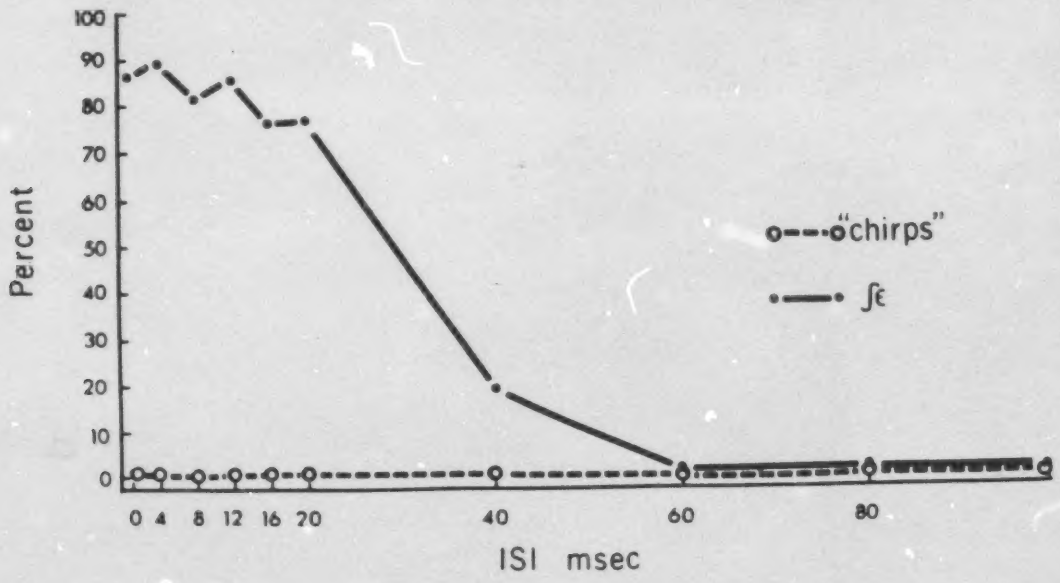


Figure 5: Percent of responses reporting the absence of the syllable-initial stop and the chirp (isolated stop cues) as a function of the interval between the end of the fricative noise and the beginning of the formant transitions.

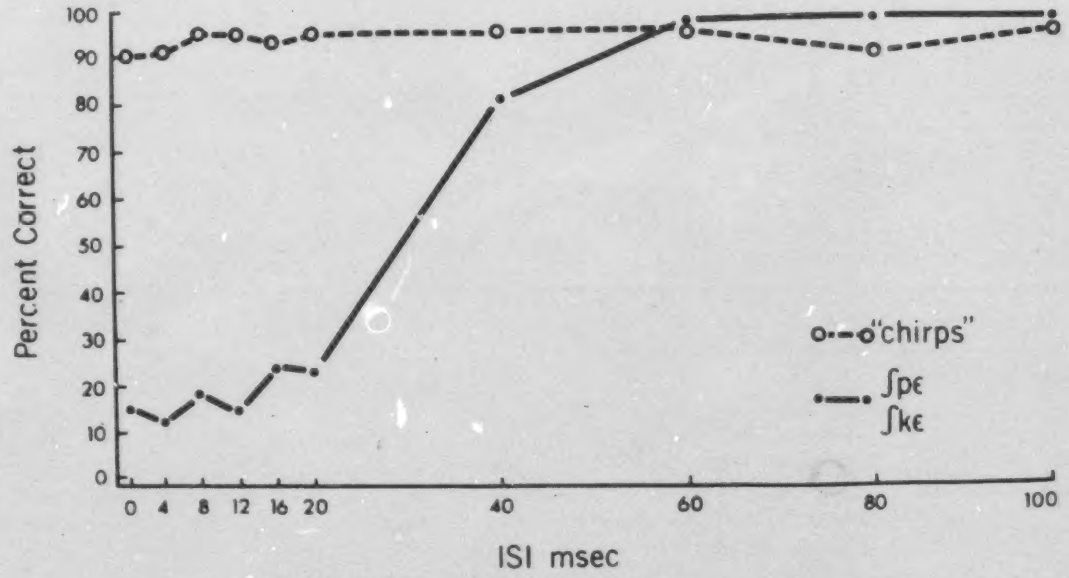


Figure 6: Percent of responses correctly identifying the syllable-initial stops and the corresponding chirps (isolated stop cues) as a function of the interval between the end of the fricative noise and the beginning of the transitions.

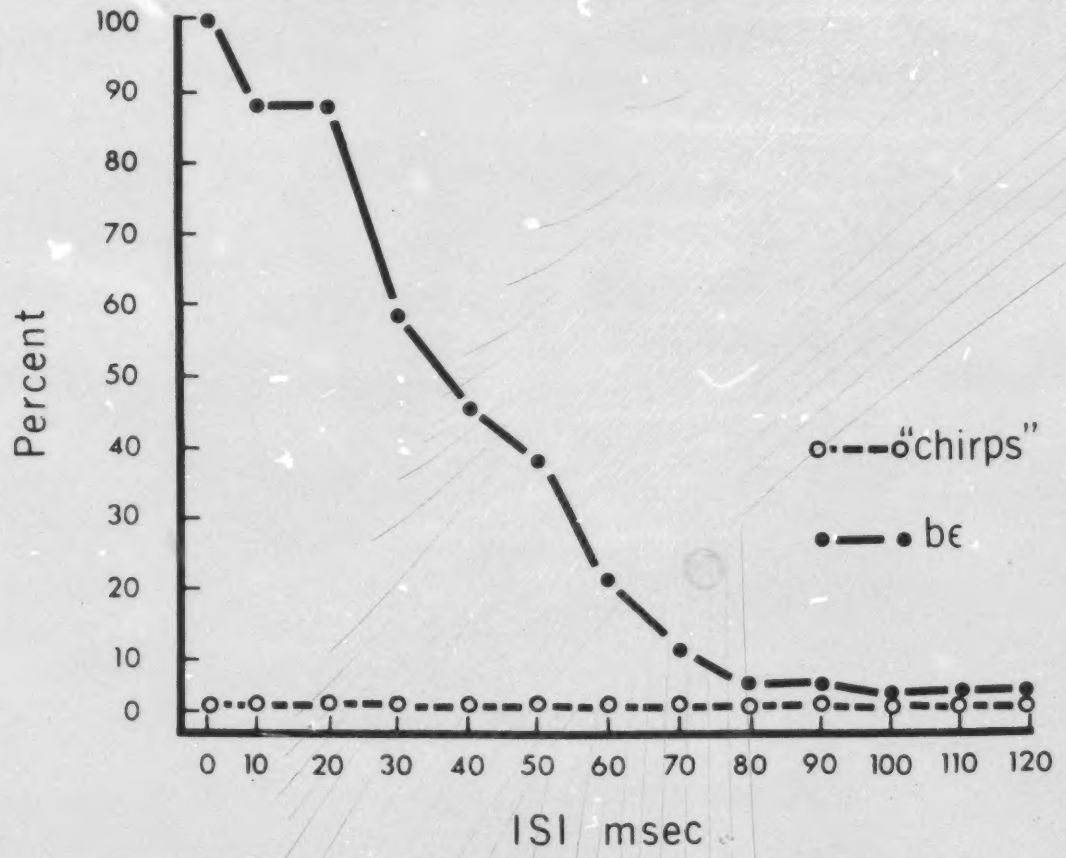


Figure 7: Percent of responses reporting the absence of the syllable-final stops and the chirps (isolated stop cues) as a function of the interval between the end of the first syllable and the beginning of the second.

Both the experiments I just described agree with the already available evidence to which I alluded and support the conclusion that there is little or no forward or backward auditory masking; more generally, they indicate that the essential stop-consonant cues are fully effective as purely auditory events, even at the very shortest intervals of silence. We should suppose, then, that the effect of the silence cue is at some other-than-auditory level, and is to be accounted for by some other-than-auditory process.

Where, now, do we look for a proper account? You students of phonetics will have been wondering to yourselves all this time why I don't take into account that a speaker cannot produce a stop without closing his vocal tract, and that the resulting silence provides information, not time to evade masking. I do want to take account of that now and to make explicit that the essential information provided to the listener is information about what the speaker's vocal tract is doing. But I would also emphasize that in this case the listener cannot perceive what the vocal tract cannot do. To see how interesting that is, contrast it with what happens in visual perception. Imagine that I show you a picture of a horse standing on its tail and ask you to tell me what you see. You would say that you see a horse standing on its tail, and you might then add

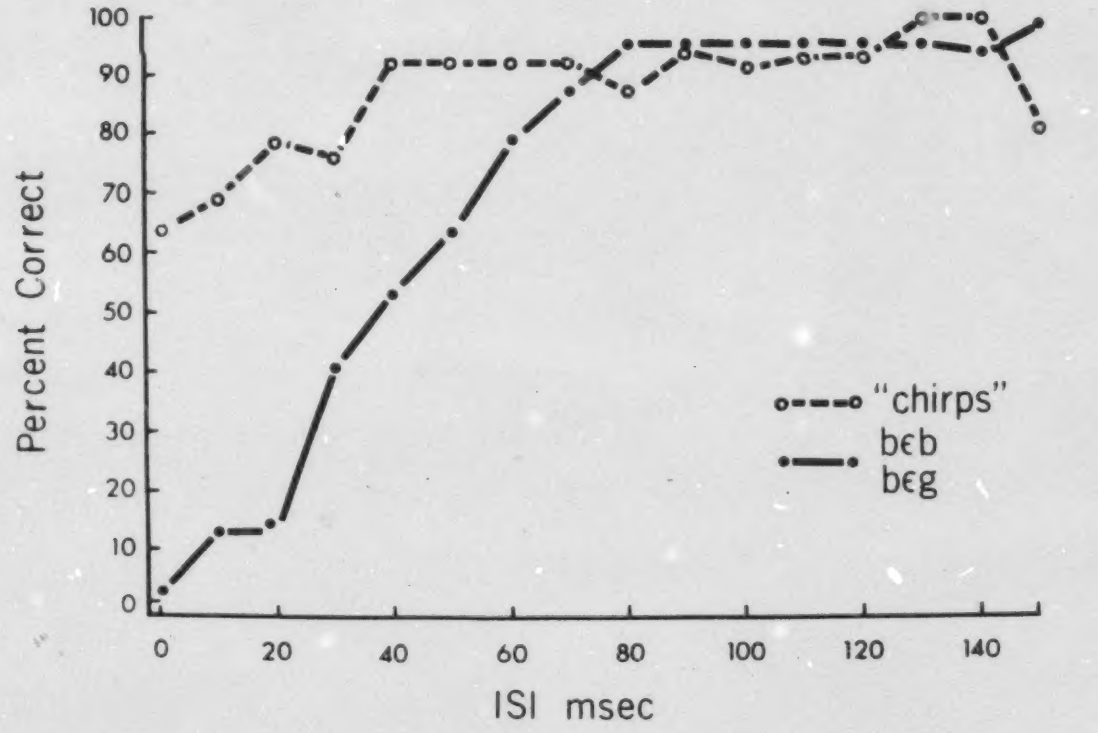


Figure 8: Percent of responses correctly identifying the syllable-final stops and the corresponding chirps (isolated stop cues) as a function of the interval between the end of the fricative noise and the beginning of the transitions.

that I had contrived the picture, since you know that horses cannot stand on their tails. But you would nevertheless have seen the horse standing on its tail. Consider how different were the results of our experiments. Had you been one of our listeners, what might you have heard and what might you have thought? Would you have heard the stop consonant and then supposed that I had synthesized the sound because you know that a vocal tract could not have responded that fast? No. You would not have heard the stop, period. Which brings us, then, to the third question relevant to our experiments: Are we here dealing with cases in which the constraint on perception is not so much by the ear as by the vocal tract?

The relevant experiments are like the earlier ones on the perception of stops in syllable-final position, except that now we use all three stops at the end of the first syllable, followed in all possible combinations by all three stops at the beginning of the next. That is, we now have [bɑb], [bɑd], and [bɑg], followed in all pairings by [bɑ], [dɑ], and [gɑ] and, as before, we vary the intersyllable interval and ask our listeners to identify the stop at the end of the first syllable.

Figure 9 shows the results. I invite your attention to the fact that the nine curves form three clusters. Let us look first at the cluster that rises most slowly, the one described by short dashes connecting solid circles. Those

curves have in common that they represent data obtained with the geminates: [bɑb bɑ], [bɑd dɑ], and [bɑg gɑ]. The production of such geminates requires a long interval of silence between the syllables; it is of interest, surely, that their perception requires a long interval too. In any event, those perceptual results group themselves most obviously according to an articulatory criterion, not an acoustic one. The same appears to be true of the other two clusters. Thus, the middle cluster--the three curves formed by the solid line connecting stars--represents those cases in which the place-of-closure for the syllable-initial stop--that is, those cases in which the place-of-closure moved from front to back: [bɑb dɑ], [bɑb gɑ], and [bɑd gɑ]. The cluster of curves that rise most rapidly--those formed by dashed lines connecting open circles--have in common that the place of closure moves, as in a hinge, from back to front: [bɑg bɑ], [bɑg dɑ], and [bɑd bɑ]. I am not prepared to say what vocal tracts do differently regarding shifts from front to back and back to front. And I will have more faith in the generality of our results--that is, more faith that there is no simple auditory basis for the clustering--if, with a variety of vowels and hence a variety of consonant cues, we nevertheless get the same result. But, taking the data as we so far find them, I believe they do suggest that perception was here constrained by properties that belong not so much to ears as to vocal tracts.

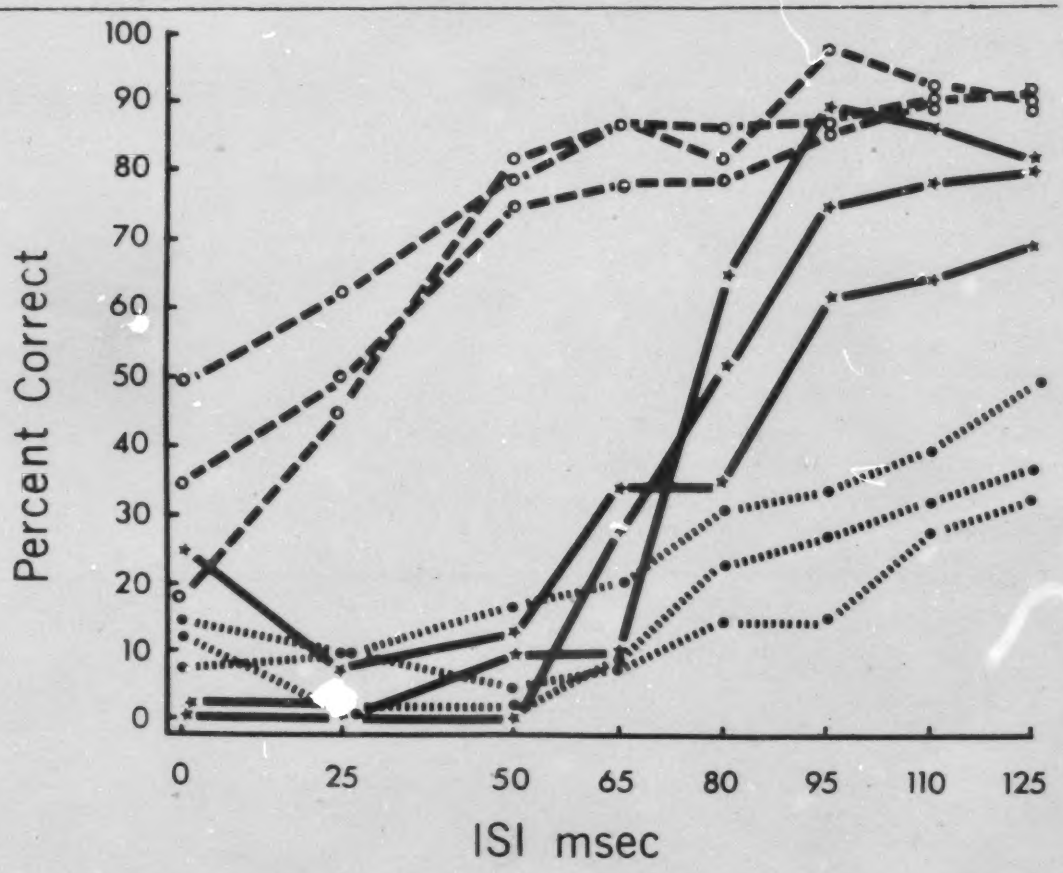


Figure 9: Percent of responses correctly identifying the syllable-final stops for all following syllable-initial stops, as a function of the interval between the end of the first syllable and the beginning of the second.

So we come now to the last question: Whose vocal tract? I should suppose that it could hardly be that of the listener or the speaker or, indeed, of any particular person, but that it must rather be an abstract conception. In this regard our situation may be similar, in its own small way, to the one in which the philosopher, Bishop Berkeley, found himself. Believing as he did that things exist only if they are perceived, he had to account for the tree that is for some time not in the mind of the gardener or of any other mortal being. He asserted that the tree did exist nevertheless because it was at all times perceived by the infinite mind of God (Berkeley, [1713] 1954). Lacking Berkeley's ecclesiastical credentials, I hesitate to make the vocal tract we are concerned with abstract in the same way. Moreover, I am not a philosopher but an experimentalist, so I must test an abstract assumption by concrete means.

We can see by experiment how abstract the theoretically relevant vocal tract must be by exploiting two facts about the ecology of speech. The first, which we have already noted, is that a speaker cannot articulate a disyllable like [bɑb dɑ] without closing his vocal tract so as to produce a silent interval between the syllables; the second is that there need be no such silent interval if the syllables are articulated by two different speakers, the first syllable by one speaker, the second by the other. Taking advantage of those facts, my colleagues and I first replicated one of our earlier experiments and found, as we had before, that when the two syllables are produced by a single speaker, the syllable-final stop can be heard only when there is a silent interval of some length between the syllables. Figure 10 shows the result. There, in the solid line connecting the solid circles, we see that, as in the earlier experiment, the listeners correctly perceived a syllable-final stop only when there was a decent interval following it. But that, I would emphasize, is what happened when the two syllables were produced by the same voice, in this case a male.

What happens when the first syllable is produced by that male voice and the second syllable by a different-sounding female voice? The answer given by eight of our first ten subjects is shown in the curve that runs almost straight across the top of the graph, the dashed line connecting open circles. That curve says that in the different-voice condition the syllable-final stop was perceived almost perfectly at all values of intersyllable intervals including even the zero value. The remaining two subjects gave results shown by the broken line connecting the open squares. As you see, they performed in the different-voice condition much the way they (and everyone else) did in the same-voice condition. It is of interest that one of those two remarked spontaneously after the experiment that she thought the voices were not different. I should add that further research on this problem has yielded results like those produced by the eight subjects; they reinforce the conclusion that in the different-voice condition the syllable-final stop can be heard at all intersyllable intervals. Thus, when the two syllables are spoken by a single voice, the syllable-final stops cannot be perceived at the short intersyllable intervals that the single voice cannot produce; but when the syllables are spoken by two different voices, production and perception are possible at all intervals between the syllables, even at no interval at all.

To develop the implications of this last finding will require many more experiments, some of which we do not yet know how to do. But one implication is fairly clear, and it happens to be one that is most relevant to our concerns in this paper. It is, as I have already said, that a motor theory must be quite

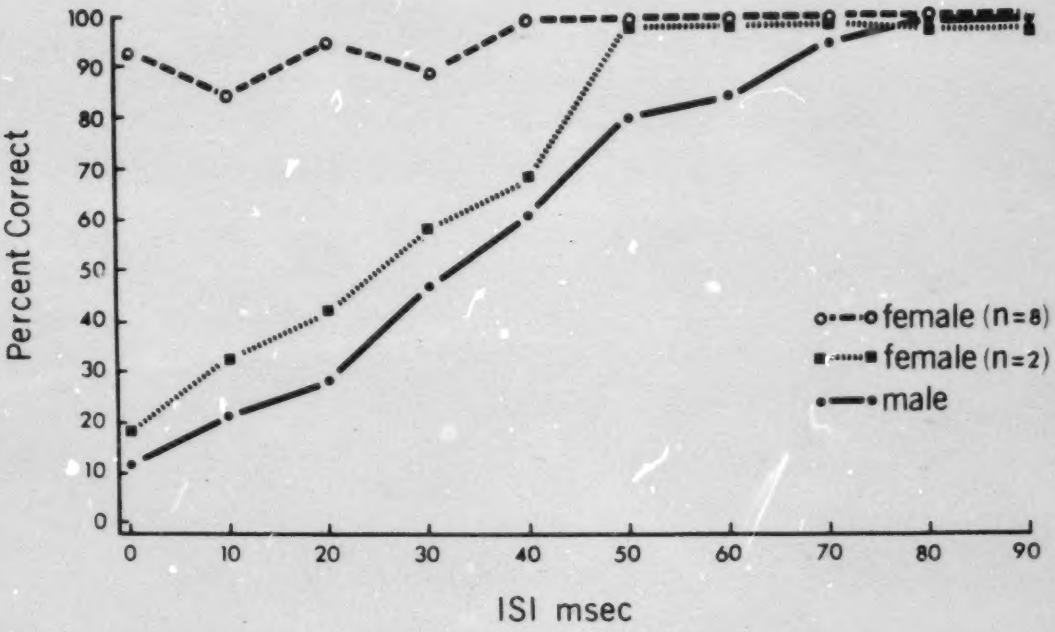


Figure 10: Percent of responses correctly identifying the syllable-final stops when both syllables are produced by a male speaker and when one is produced by a male, the other by a female.

abstract. But if our results require that the theory be so qualified, they do not yet suggest that it be wholly abandoned. To take account of these results I should suppose that the articulatory model the listener has access to acts as if it knew about the capabilities of vocal tracts in general. Computations carried out in terms of that model might yield something like the results we have observed.

I should summarize.

By examining the function of grammatical codes, we see why specialized linguistic processes might be necessary and why perception and production might be linked. In the case of speech we can be more explicit about the need for grammatical recoding and especially about the fit of grammatical form to perceptual function. We see more clearly, then, that specialized perceptual processes might be required if the listener is to cope with the code, and we see just as clearly that the key to the code is in the manner of its production. Because we suspect that such a key may be a part of the specialized processes, we look with favor on a motor theory of speech perception.

Several recent experiments illustrate that data relevant to such a theory can be obtained. In these experiments it was found that (1) silence is an important manner cue for the perception of stop consonants; (2) this cue is not constrained primarily by the properties of the auditory system; (3) the constraint

appears rather to be related to what the vocal tract can and cannot do; and (4) it is not any particular vocal tract that imposes the constraint but some conception of vocal tracts in general. I should conclude that a motor theory is still a reasonable way to make sense of some of the phenomena of speech perception, but only if we assume that the implied reference to production is highly abstract.

REFERENCES

Berkeley, G. (1954) Three Dialogues between Hylas and Philonous. (New York: Library of Liberal Arts).

Darwin, C. J. (1969) Auditory perception and cerebral dominance. Unpublished Ph.D. dissertation, University of Cambridge.

Dorman, M. F., L. J. Raphael, A. M. Liberman, and B. Repp. (1975) Maskinglike phenomena in speech perception. J. Acoust. Soc. Am. 57, Suppl. 1, 48(A). [Also in Haskins Laboratories Status Report on Speech Research SR-42/43 (1975), 265-276.]

Elliott, L. L. (1971) Backward and forward masking. Audiology 10, 65-76.

Ganong, W. F. (1975) An experiment on "phonetic adaptation." Progress Report (Research Laboratory of Electronics, MIT) PR-116, 206-210.

Leshowitz, B. and E. Cudahy. (1973) Frequency discrimination in the presence of another tone. J. Acoust. Soc. Am. 54, 882-887.

Liberman, A. M. (1974) The specialization of the language hemisphere. In The Neurosciences: Third Study Program, ed. by F. O. Schmitt and F. G. Worden. (Cambridge, Mass.: MIT Press), pp. 43-56.

Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1973) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington, D.C.: V. H. Winston), pp. 307-334.

A Systems Approach to the Cerebral Hemispheres

Carol A. Fowler*

ABSTRACT

There is evidence that the two cerebral hemispheres are at birth equally capable of acquiring language. There is also evidence that control over linguistic function becomes, in due course, the domain of only one hemisphere, usually the left. We are thus confronted with the paradox that a neurological system (the right hemisphere), though fully equipped to control a particular function, normally develops in such a way that it fails to do so. Cerebral dominance theory has recently been revised by several theorists, partly in order to resolve this discrepancy between the linguistic potential and the linguistic achievement of the right hemisphere under normal conditions of development. The resolution is affected by expanding the concept of cerebral dominance to include the notion of active control, through inhibition, of one hemisphere by the other. The language-dominant left hemisphere is thus considered to inhibit the right. However, rather than resolve a paradox by modifying theory, a more satisfactory solution is to remove it. This may be accomplished by adopting a novel perspective on the apparently discrepant observations, so that they lose their paradoxical appearance. The present paper argues that such a perspective is provided by a theory of dominance modeled after the tenets of General Systems Theory.

The concept of cerebral dominance has been radically revised in the last several years. In 1962, Zangwill described the then accepted view of dominance as the asymmetric representation of function in the human cerebral hemispheres. This view might be called "static," since it merely locates a particular function in a particular hemisphere and says nothing about the functional relations between hemispheres. The static view may be contrasted with a second perspective on dominance that conforms more to the usual meaning of the word. Dominance, according to the second view, refers to the active control of one hemisphere by the other. Zangwill considered this dynamic interpretation, but rejected it on grounds of parsimony: although the dynamic view was compatible with the static, he saw nothing in the available evidence to demand it.

---

*Also University of Connecticut, Storrs.

Acknowledgment: I wish to express my appreciation to Michael Turvey and Michael Studdert-Kennedy for reading and critiquing earlier versions of this manuscript.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

Recently, several investigators have seen the matter differently and have therefore incorporated the dynamic perspective into their accounts of cerebral dominance (Geschwind, 1969; Gazzaniga, 1970; Gazzaniga and Hillyard, 1973; Kinsbourne, 1973, 1974; Moscovitch, 1973; Selnes, 1974). One impetus for augmenting the concept of dominance was the recognition of a paradox deriving from the following observations. First, there is convincing evidence that the human cerebral hemispheres at birth are more or less equipotential in their capacity to become the language dominant hemisphere (Lenneberg, 1967). Second, the right hemisphere only realizes its potential under unusual conditions. The paradox, then, is that a "language center" develops in only one hemisphere, even though the capacity to develop language is present in both. Recent qualifications of the equipotentiality of the hemispheres--provided by evidence that the left hemisphere is both structurally (Witelson and Pallie, 1973) and functionally (Molfese, 1972; Entus, 1975) specialized before language acquisition begins--do not resolve the paradox, since the fact remains that, under conditions of early left-hemisphere removal, the right hemisphere does acquire language.

Indeed, the paradox cannot be resolved within the static dominance scheme, for it holds that dominance is total: a particular function is the exclusive domain of a particular hemisphere. Hemispheric specialization is thus a fait accompli; there is no provision within the scheme for the gradual establishment of asymmetric representation in initially equipotential hemispheres. Therefore, the evidence that hemispheric representation of function is not deterministically asymmetric from birth demands either a modification of the static view or its rejection.

There are two ways in which a scientific paradox may be handled. Theory may be modified expressly to resolve it. Or the paradox may be dissolved by adopting a new perspective from which the discrepant observations lose their paradoxical appearance. The first tack has been taken in two recent accounts of development of cerebral dominance, namely those of Selnes (1974) and of Gazzaniga (1970). Both investigators modify the static view by incorporating within it the dynamic notion of dominance as control over one hemisphere by the other. The second tack will be taken here. We shall argue that neither Selnes nor Gazzaniga provides a satisfactory account. More importantly, we shall argue that total rejection and replacement of the static view with a Systems Theoretical view of dominance development provides a simpler solution to the problem. Before characterizing these two attempted resolutions of the paradox, and their consequences for the theory of hemispheric specialization, let us lay out in more detail the specific observations that require explanation.

HEMISPHERIC EQUIPOTENTIALITY IN EARLY CHILDHOOD

The literature on cerebral dominance provides information about language acquisition and hemispheric specialization for language in a variety of neural contexts. Studies of language acquisition in cases of callosal agenesis and of early hemispherectomy reveal the linguistic capacities of the isolated left and right hemispheres. These capacities may be compared with those of left and right hemispheres that were connected during language acquisition. The latter evidence comes from normal individuals and from individuals who underwent commissurotomy or hemispherectomy as adults.

Differences in the neural contexts of language acquisition give rise to apparent differences in the locus of language representation. In no case does

language acquisition fail to occur. In the normal individual, language function comes to be controlled primarily by a single hemisphere, usually the left, while the right hemisphere demonstrates very limited linguistic skills if examined during left-hemisphere anesthetization (Milner, Branch, and Rasmussen, 1964), after commissurotomy (see, for example, Gazzaniga, 1970) or after hemispherectomy (Smith and Burklund, 1966). We might therefore adopt an extreme view of dominance, and say that language is normally represented in the left hemisphere, but not in the right.

However, if one hemisphere is removed early in life, language develops in the remaining hemisphere regardless of which has been removed (see, for example, Basser, 1962). Thus if the left hemisphere is removed, language is acquired by the remaining right hemisphere. In cases of callosal agenesis, the evidence is less clear, though not incompatible with the interpretation that "language centers" are established in both hemispheres (Bryden and Zurif, 1970; Sperry, 1970; Saul and Scott, 1973).

In any event, we have clear evidence that the right hemisphere has the capacity to acquire and represent language, and the paradox is that this capacity is not realized under normal conditions of development. The resolutions of the paradox offered by Gazzaniga (1970) and by Selnes (1974) take as their starting point the above descriptions of the conditions under which right-hemisphere language acquisition occurs or fails to occur. The systems approach advocated in the present paper will argue that the conditions are both inappropriately characterized and inappropriately interpreted by the proposed resolution.

MODIFICATIONS OF THE STATIC DOMINANCE VIEW: PARADOX RESOLUTION

The observations of the preceding section suggest an obvious solution to the paradox. Since the right hemisphere, under normal developmental conditions, fails to realize its demonstrated potential to acquire language, something must prevent it from doing so. The nature of the conditions under which it does realize its language acquisition potential--for example, early left hemispherectomy--suggests that the source of right-hemisphere suppression lies in the left hemisphere. This is the reasoning of Selnes and of Gazzaniga. The mechanism of right-hemisphere suppression they both propose is inhibition, initiated by the left hemisphere and mediated by the corpus callosum (Gazzaniga, 1970; Gazzaniga and Hillyard, 1973; Selnes, 1974).

The inhibition hypotheses proposed by Gazzaniga and Selnes are not identical; in fact they are complementary. Gazzaniga (1970), elaborating on Hewitt's (1962) evidence that the corpus callosum is not fully developed at birth, argues that the infant is functionally split-brained for the first two years of its life. During this period, before the onset of suppression, the right hemisphere acquires those minimal linguistic skills that can be demonstrated in the isolated adult right hemisphere. As the corpus callosum matures, the left hemisphere begins to suppress the right, and further right-hemisphere language acquisition is prevented. For Gazzaniga then, the term dominance has a dual reference: it refers both to the permanent active control of one hemisphere by another, and simultaneously to the consequence of that control--the asymmetric representation of function in the hemispheres.

Selnes (1974), on the other hand, recognizes some of the fairly strong evidence against the inhibition hypothesis, at least as applied to adult hemispheres

(see next section). Therefore, since the facts of dominance development appear to demand an inhibition interpretation, he proposes that the period of right-hemisphere inhibition by the left is restricted to infancy and early childhood. For Selnes, that is, dominance has two phases: an initial dynamic phase of active left-hemisphere control of the right, and a subsequent static phase of dominance as asymmetric respesentation of function.

The development of asymmetric representation in the initially equipotential hemispheres can now be explained. Both hemispheres have the potential to acquire language. But the left hemisphere is additionally equipped to inhibit the right, to prevent it from acquiring language during infancy, and, according to Gazzaniga (1970), to prevent the interference of any primitive right-hemisphere linguistic activity with left-hemisphere processing in the mature brain. However, if the source of inhibition is removed while the brain is still immature, or if the pathway mediating the inhibitory influence fails to develop, the right hemisphere's capacities are realized.

In sum, both the above resolutions of the dominance development paradox involve the addition of a very powerful construct--that of active inhibition--to the theory of cerebral dominance. As we have seen, this construct appears to be required by the facts of cerebral dominance development. Moreover, its addition can be justified on grounds independent of those facts (Geschwind, 1969; Kinsbourne, 1970, 1973; Gazzaniga and Hillyard, 1973; Moscovitch, 1973). However, strong objections can be raised to the inhibition hypothesis, even within the static dominance view, and some of these will now be considered.

INHIBITION HYPOTHESES

The inhibition hypothesis takes several different forms, each largely shaped by the facts it is supposed to explain. First, both Kinsbourne (1973) and Gazzaniga and Hillyard (1973) have postulated a bidirectional reciprocal inhibitory relation between the hemispheres. Kinsbourne does so in order to complement a proposed mechanism for focusing attention on operations taking place in one or the other hemisphere; Gazzaniga and Hillyard (1973), in order to account for the observed increase in total processing capacity of the brain following commissurotomy. Second, Geschwind (1969) and Moscovitch (1973) have argued for a unidirectional, focalized, and tonic inhibition of right-hemisphere language centers by language centers in the left hemisphere. The grounds for this hypothesis are comparisons of right-hemisphere linguistic abilities among normals, commissurotomees, adult and child aphasics, and left-hemispherectomized individuals. Finally, as we have seen, Selnes (1974) has proposed unidirectional inhibition of the right hemisphere by the left during infancy, largely to resolve the developmental paradox described above.

We will consider each of these hypotheses in turn.

Inhibition as Reciprocal Inhibition

Gazzaniga and Hillyard (1973) offer the following description of the proposed interhemispheric inhibitory operation:

> ...the role of the forebrain commissures in integrating the attentional processes of the two cerebral hemispheres is revealed by the increases of total processing capacity upon the removal of the corpus

> callosum. It is as if in the normally interconnected brain the callosum is involved in inhibiting the transmission of information undergoing processing extraneous to the dominant cognitive activity under consideration. The brain cannot consider all things at all times, and perhaps order only is brought about by what amounts to a cognitive counterpart of a reciprocal inhibition kind of mechanism. (p. 237)

Kinsbourne's (1970) proposal is similar in form although he derives it from a different set of observations. He describes his attentional model as follows:

> Each hemisphere serves the contralateral half of space.... Thus, as a matter of course, orientation to one side of space coincided with preparatory activation within the contralateral hemisphere. If the principle of reciprocal innervation holds not only at spinal cord level (Sherrington, 1906),[1] but also between the cerebral hemispheres (Kinsbourne, 1970),[2] then as one hemisphere actively subserves its orienting function, the other is inhibited as regards the contrary tendency it subserves. (pp. 195-196)

There are five main objections to the reciprocal inhibition hypothesis.

(1) The analogy made by Kinsbourne (1973) and by Gazzaniga and Hillyard (1973) between spinal reciprocal inhibition and the proposed inhibitory effect is not a close one. In the spinal cord, reciprocal inhibition works to offset the tendency of a muscle antagonist to counteract agonist activity. When an agonist contracts, antagonist muscle spindles are stretched and their spindle afferents are excited. If uninhibited, the spindle afferents elicit (via their connections to the α motorneurons innervating the antagonist) antagonist contraction that counteracts the effect of agonist activity. Therefore, if voluntary movements are to occur, antagonists must be prevented from counteracting agonist activity, and this appears to be the role of spinal reciprocal inhibition.

Notice that the neural system innervating an agonist does not inhibit the motorneurons of every muscle that might interfere if activated simultaneously. Rather it specifically inhibits those motorneurons innervating the muscle that is structurally designed to counteract its effects. Unless we assume that, like agonist and antagonist, the left and right hemispheres are designed so that activities in one hemisphere are counteracted by subsequent activities in the other, the reciprocal inhibition analogy would seem to be inappropriate to a description of interhemispheric relations.

(2) The reciprocal inhibition hypothesis is inconsistent with data demonstrating a much greater degree of suppression in commissurotomized individuals than in normals. These data suggest that the normal corpus callosum may mediate

---

[1] Sherrington, C. S. (1906) The Integrative Action of the Nervous System. (New York: Scribner).

[2] Kinsbourne, M. (1970) The cerebral basis of lateral asymmetries in attention. Acta Psychologica 33, 193-201.

arousal rather than suppression. The split-brain suppression effect is reported quite frequently in the literature, and three examples follow.

(i) Trevarthen (1970) reports that callosalized human subjects, asked to fixate the center of a table on which an irregular shape, cut out of white card, had been placed and to mark the center of the card, experienced a fading or disappearance of the shape, if it was located in the visual field contralateral to the responding hand. If the responding hand was the left hand, and the shape was in the right visual field, the subject reported that he could not see the object. Trevarthen notes that this perceptual "neglect" was not typically observed when the required response was less skilled and more automatic than the marking response.

(ii) Trevarthen and Sperry (1973) report similar perceptual neglect effects among callosalized subjects who were asked to compare stimuli presented in different visual fields. Unilateral neglect occurred when the subject was asked to describe the stimuli or when manual responses were required. For example, subjects often neglected the left visual field stimuli when responding verbally, or, if they tried to express what they had apparently perceived, suffered an arrest of speech. In at least one subject, unilateral neglect declined in frequency as a session progressed--as the subject, according to Trevarthen and Sperry, "developed sufficient concentration of his attention on the task."

(iii) Teng and Sperry (1973, 1974) report that under conditions in which digits or dots were presented for identification in both visual fields (subjects held out a number of fingers corresponding to the digit or to the number of dots displayed), callosalized subjects tended to neglect one visual field. Neglect was not observed under conditions of unilateral presentation of digits or dots.

Since these suppression effects are peculiar to individuals lacking a corpus callosum, it is unlikely that the role of the callosum in normal individuals is to suppress functioning. The evidence suggests rather that it may mediate arousal.

(3) Callosal anatomy and physiology suggest that the callosum functions to permit communication between split sensory fields rather than to suppress functioning. Selnes (1974) reviews literature showing that callosal fibers between corresponding projection areas in the hemispheres only connect those parts of the projection areas representing the midlines of the sensory fields. There are no striate-striate callosal connections in the visual system. Callosal fibers only connect the extrastriate areas, which represent the midline of the visual field. Similarly in the somesthetic and motor areas, fibers interconnect primary projection areas representing axial body structures, but not the hand, finger, or foot areas. "Association areas" show a similar duality--that is, well-defined areas that are interconnected across the callosum and others that are not. At least with respect to the projection areas, the specificity of callosal connections to areas representing sensory or motor midlines suggests that the role of the callosum is to permit communication between areas of the brain representing neighboring areas of space that are split between the hemispheres.

(4) Measurements of left- and right-hemisphere-evoked potentials during the performance of linguistic and nonlinguistic tasks provide evidence incompatible with reciprocal interhemispheric inhibition. Wood, Goff, and Day (1971) found that the right hemisphere responded identically to syllables presented auditorily,

regardless of the linguistic or nonlinguistic nature of the subject's task. The left-hemisphere-evoked potential, on the other hand, varied with the nature of the task. Although it is not clear what the evoked potential represents, one might reasonably expect that, if the right hemisphere was inhibited in one case and not in the other, its evoked potential would vary accordingly.

(5) A final objection to the reciprocal inhibition hypothesis is simply that it is unnecessary. We do not need it to explain either the failure of the hemispheres to interfere with one another under normal conditions or the occurrence of interference effects during the simultaneous performance of two tasks by the hemispheres (Geffen, Bradshaw, and Nettleton, 1973; Hicks, 1975; Kinsbourne and Cook, 1971). These effects can be satisfactorily accounted for by attentional mechanisms similar to those proposed by Kinsbourne (1973) and by Trevarthen (1974) to explain perceptual neglect effects among callosalized subjects.

Inhibition as Focalized Unilateral Suppression of Right-Hemisphere Language Centers

Most of the arguments marshaled against the notion that the corpus callosum mediates reciprocal inhibition apply, of course, to this inhibition hypothesis as well. However, Moscovitch (1973) suggests that some part of the inhibition effect must be assumed to be subcortical if the continued inability of the right hemisphere to initiate speech following commissurotomy is to be explained. Consequently, the focalized inhibition hypothesis cannot be rejected on the grounds that inhibition is not mediated callosally.

The observations explained by an inhibition hypothesis are, according to Moscovitch:

1. the apparent inability of the right hemisphere in normal adults to process verbal information,

2. improved right-hemisphere linguistic abilities following commissurotomy,

3. the inferior right-hemisphere linguistic abilities of some aphasics relative to split-brain subjects.

There appear to be two grounds for the claim that the normal right hemisphere cannot process verbal information. First, individuals are typically unable to initiate speech during left-hemisphere anesthetization. We should note, however, that according to Milner, Branch, and Rasmussen (1964) this effect gives way, after a few minutes, to a transient aphasia. If the aphasia is symptomatic of left-hemisphere recovery from the anesthesia, then the claim that the right hemisphere is mute is based merely on its performance in the very few minutes before left-hemisphere recovery. We cannot justifiably conclude from this that the right hemisphere's muteness would persist indefinitely if the period of left-hemisphere suppression were extended.

The second ground for Moscovitch's assessment of normal right-hemisphere linguistic abilities comes from his own work on visual field effects (Moscovitch, 1973). Subjects in these experiments listened binaurally over earphones to

either one or six letter names (such as "bee," "cee," "dee," etc.) and were then presented with a letter in either right or left visual field. Subjects indicated whether or not the visually presented letter had been among those presented auditorily. A left visual field reaction time advantage was obtained for the single letter memory set, while a right visual field advantage was obtained for the six item set. When the single letter memory set condition was made "more linguistic" by requiring subjects to respond to letters that were rhymes of the auditorily presented letter as well as to the letter itself, a right visual field advantage was obtained. The evidence indicated to Moscovitch that the right hemisphere among normals is unable to process verbal information.

He then argues that, since right-hemisphere speech perception has been demonstrated among split-brain subjects, these speech abilities must be inhibited in the normal right hemisphere. We should note, however, that the stimuli used by Moscovitch were letter names, while those used to test callosalized subjects had, until recently, been real words. Consonant-vowel (CV) nonsense syllables were used, both monaurally and dichotically, by Zaidel (1974) to test right-hemisphere perception in split-brain subjects. He found that the subjects were unable to identify the syllables presented to the right hemisphere by pointing to their letter representations. Thus, split-brain subjects fail to demonstrate linguistic skills superior to those of the normal right hemisphere when the stimuli are nonsense syllables, and Moscovitch may have unwittingly chosen the wrong stimuli to test his hypothesis. On the other hand, Dimond (1971) has shown that visually presented letter sets are better reported by normals if the letter sets are divided over both hemispheres than if they are sent to a single hemisphere. This indicates that some part of the letter sets was processed by the right hemisphere. Hence Moscovitch's claim that the normal right hemisphere cannot process verbal stimuli may be incorrect. In fact, his reaction time technique may not establish which hemisphere is uniquely able to process verbal stimuli, but merely which hemisphere processes them faster.

Moscovitch's first two claims, therefore, that the normal right hemisphere is unable to process verbal information and that the right hemispheres of callosalized individuals exhibit improved linguistic abilities relative to the normal right hemisphere, may be unjustified. His final claim, that some aphasics demonstrate less comprehension of speech than does the right hemisphere of callosalized individuals, may well be true. However, as we shall see below, it may be attributed as readily to interference by a malfunctioning left hemisphere as to inhibition.

Inhibition in Infancy

Selnes's (1974) proposal that inhibition of the right hemisphere by the left is restricted to infancy purports to explain first, as noted above, the more or less unilateral development of language despite apparent left- and right-hemisphere equipotentiality during infancy, and second, the rapid improvement of linguistic skills following dominant hemispherectomy for infantile hemiplegia (see, for example, Basser, 1962).

Although inhibition of the right hemisphere by the left (and its release following left hemispherectomy) could account for these phenomena, inhibition cannot be the simplest explanation, nor is it a very likely one. We have already reviewed evidence suggesting that the role of the corpus callosum is to transmit

information between the hemispheres, and Selnes (1974) himself rejects the notion that the corpus callosum mediates inhibition in the mature brain. The hypothesis that it mediates inhibition in infancy therefore requires the unlikely assumption that there exists some inhibition mechanism that disappears or changes its character as the organism matures.

Beyond this, and more generally, the reasoning of Selnes's proposal [like that of Gazzaniga (1970)] is specious. Briefly, the reasoning was this: if the right hemisphere has the potential to acquire language, but only does so when it is isolated from the left, then the left hemisphere must normally prevent the right from developing its potential. A plausible neurological mechanism for this process is inhibition.

But does logic compel us to conclude that the left hemisphere "prevents" the right from developing, and how plausible, in fact, is the inhibition hypothesis? A radio damaged by a hammer blow may emit a continuous howl; after a stroke, a person may walk with a limp or may only be able to produce jargon when he tries to speak. Yet we would not be inclined to claim that the novel behaviors following injury were "released" because their inhibitors were damaged (cf. Gregory, 1961). The radio does not emit a howl because its howl inhibitor was damaged by the hammer blow; nor does the aphasic produce jargon because his jargon inhibitor has been destroyed. The novel responses arise because the injured system--the radio or the brain--is different both structurally and functionally from the system that it was before injury. Some of its components have been destroyed, and the interrelations among the remaining intact components have been altered as a consequence.

Of course, the right hemisphere's "response" to isolation is not entirely comparable to the radio's howl or even to the aphasic's jargon. Whereas the latter are pathological and maladaptive, the acquisition of language is highly adaptive. Nonetheless, we may account for right-hemisphere language acquisition after hemispherectomy in the same general way that we account for the emergence of the howl and the jargon following accident or injury. In the intact brain the two hemispheres comprise a single system of interdependent components. When a subset of these components is removed, the functioning of the rest changes in consequence. By this account, the isolated right hemisphere is not functionally the same system as the intact connected right hemisphere.

In short, the development of language in the right hemisphere following its isolation from the left does not <u>necessarily</u> imply that it has been freed from a language-center inhibitor. There is at least one alternative view: that isolating the right hemisphere from the left effects a change in the right hemisphere's mode of functioning. A new system then emerges with the capacity to acquire language. It is this alternative view that we consider in the following sections, and that we will elaborate in relation to the dominance development paradox and the theory of cerebral dominance.

ADOPTION OF THE SYSTEMS APPROACH: PARADOX DISSOLUTION

Systems Theory (Bertalanffy, 1968; Weiss, 1969, 1971) provides a perspective on hemispheric specialization from which the observations on cerebral dominance development lose their paradoxical appearance. Indeed, the observations are closely analogous to certain characteristics of developing biological

organisms that are frequently cited to illustrate fundamental systems properties. Many of these properties can be derived from the following preliminary definition of a system: a system is a whole or a unit composed of hierarchically organized and functionally highly interdependent subunits that may themselves be systems. The following are examples of systems: an atom, a cell, a person, a factory, a society. The systems-theoretical perspective on these instances of "organized complexity" reveals that they share certain fundamental properties. Moreover, despite the diversity of the organizations properly described as systems, these properties are not so general as to be trivial or useless. Some of them will be described below.

Weiss (1969, 1971) characterizes the functional interdependence of the systems subunits in terms of the following inequality: $V_s < (V_a + V_b + \ldots + V_n)$, where V, s, and a-n stand for "variance," "system," and "subunits a to n," respectively. According to the inequality, the variance in the states of the system as a whole is less than the sum of the variances of the individual subunits. To illustrate how this property manifests itself, let us look at an example of a system in which it is clearly revealed. The system is the speech production system, and we will examine its output, the spoken word. Lehiste (1971) described an experiment in which a talker is asked to repeat a word 50 times. The duration of each repetition is measured and its variance across repetitions is computed. Additionally, the durations of the component acoustic segments are measured on each repetition and variances are computed for each. The sum of the variances of the component segment durations consistently exceeds the variance of the total word duration, in Lehiste's experiment by a factor of 3 to 5. What this implies, as Lehiste points out, is that on a repetition in which some segments are unusually long in duration the others must be correspondingly short. That is, in order for the variance of a whole to be small relative to the sum of the variances of its parts, the parts must be compensating for deviations in each other's behavior.

That the durations of the individual acoustic segments vary at all across repetitions of a word of relatively fixed duration indicates that they are not or cannot be rigidly controlled by the speaker. Weiss (1969) terms the corresponding systems property "microindeterminancy." Nonetheless, although the subparts are not rigidly controlled as individuals, their collective behavior is relatively controlled; the speech production system consistently reaches a (relatively) fixed goal, despite the microindeterminacy of its component subparts. More generally, a system is an organization whose overall state is stable relative to the states of its components. These observations suggest an important conclusion: temporal compensation may occur among the acoustic segments of a word because the talker establishes or plans the total word duration from the outset. The target-word-duration then exerts a regulatory influence throughout the course of each word's production. We can think of the duration specification as the fixing of a potential and crucial degree of freedom in the system--the endpoint that the system is then constrained to reach. The established endpoint exerts a regulatory influence on the subsequent behavior of the system. The influence must be regulatory rather than controlling because only the endpoint is set, not the route by which the endpoint is reached.

Generalizing the conclusion to systems of all kinds, we can say that a system whose goal state has been set is constrained to reach that state, and that the goal specification regulates the course of goal attainment. However, since

the influence is not microdeterministic, the system can reach the goal by a variety of routes. This tendency for a system to reach a constant endpoint from a variety of starting points and by a variety of routes is called "equifinality." It is characteristic of any organization that, as a whole, has functional properties that do not inhere in any of the subparts individually. (In the speech production system, for instance, word duration is not a functional property of any of the system subparts responsible for producing the different acoustic segments.) These functional properties or systems dynamics constitute an equilibrium state of the system. They define the configuration or endpoint toward which the system will tend; they do not, however, define the precise way in which it will attain that configuration.

Examples of the diverse kinds of organizations that exhibit the equifinality characteristic (and do so for the reasons described above) are the speech production system, the joint-muscle system responsible for setting a joint angle (Asatryan and Fel'dman, 1965), the developing embryo, and, we will argue, the developing cerebral hemispheres.

So far, two closely related systems properties have been described: the stability of the system as a whole relative to variations in the states of its component parts (microindeterminacy), and the regulation of changes in the system's states by means of preset goals (equifinality). A final systems characteristic can be derived from the information already compiled. We have noted the existence of functional systems properties that do not inhere in any of the subparts. The existence of such properties implies a certain degree of independence of the systems dynamics from the structural units over which they operate. This independence provides the system with a corresponding degree of resistance to destruction due to the loss of individual subunits. To cite an example given by Weiss (1971), the death of a cell does not destroy the systems dynamics in a biological organism. It does not, because the functional properties of the system do not depend for their realization on the performances of particular cells. Functions are not coded in terms of individual subunit behaviors.

Experiments performed on the developing embryo are frequently cited by systems theorists, in part, because they clearly demonstrate this last systems characteristic. Here is Weiss's (1971:22-23) description of the experiments:

> It was only consistent on the part of performationists, who adhered strictly to a machine-like concept of development, that upon seeing a <u>whole</u> embryo develop from each <u>half</u> of a bisected egg, they would presume each blastomere of the 2-cell stage to be endowed with a spare mechanism for the formation of a whole embryo, to be activated in just such an emergency as accidental blastomeral separation. What neither they nor evolution could have foreseen was that enterprising human experimenters would move on in the opposite direction and fuse two whole eggs, with the result that a single harmonious giant embryo would form from the fused mass.... Since contrary to splitting, the natural occurrence of such a merger would be impossible, among other reasons because of the barrier of the enveloping egg membranes, it would have been absurd to postulate the providential inclusion by evolution of a spare mechanism for half an embryo in a whole egg. This once and for all disposed of the notion of spare mechanisms predesigned for developmental correctives, and by the same token, also

> of wholly rigid preformed mechanisms for the normal course of development as such.
>
> We are thus compelled to fall back on pure and unreducible system behavior as an indispensable principle of developmental dynamics.

Weiss was, of course, mistaken in his claim that the experiments "once and for all disposed of the notion of spare mechanisms predesigned for developmental correctives," since that notion is manifest in the inhibition hypotheses of hemispheric equifinality described above.

There are two possible analogies between embryo development and the development of cerebral dominance for language: (1) Each half of a bisected egg develops into a whole functioning organism. Similarly, each hemisphere of an immature acallosal (bisected) brain develops a language function, as does the remaining half-brain after early left or right hemispherectomy. (2) Two fused eggs develop into a single organism. Analogously, the two normally connected hemispheres develop a single language function. The similarities between the two sets of observations suggest both that the "spare mechanism" account of right-hemisphere language potential is not correct, and that the hemispheres might profitably be viewed as instances of a system.

The hemispherectomy and callosal agenesis data constitute a demonstration of the systems properties of equifinality and of the partial independence of systems dynamics from the structural units over which they operate. Because, on the systems view, the developmental sequence leading to dominance is not coded in terms of particular kinds of changes in particular systems subunits (because it is, in fact, a set of possible sequences sharing a common endpoint), it is resistant to destruction due to the loss of individual subunits. It is apparently resistant even to the loss of an entire cerebral hemisphere. Hence, when a hemisphere is removed, or when the fiber tract connecting the hemispheres fails to develop, the systems dynamics remain intact. However, the domain over which they operate becomes a single hemisphere instead of two. (Two independent sets of systems dynamics are considered to operate in the acallosal brain.) The endpoint or goal that the dynamics define is attained in the remaining whole.

Clearly, if we adopt the systems view, interhemispheric inhibition is not required to "prevent" the right hemisphere from realizing its potential to acquire language. Indeed, the right hemisphere has no "spare mechanism" for language development that might require inhibition. The development of dominance is not the realization of some potential inherent in the individual subunits themselves. Rather it is the product of systems dynamics that, in the normal hemispheres, regulate the whole brain. Nor are the hemispheres considered to be independent units competing for functional prepotency; they are interdependent systems subunits tending toward the same goal or equilibrium state.

There is a final case in the hemispheres literature that has no analogue in the series of experiments briefly described by Weiss (1971). Language functions in the adult or child may be impaired more seriously before than after surgical removal of a damaged left hemisphere. This observation is considered by some theorists (see, for example, Geschwind, 1969; Moscovitch, 1973) to constitute evidence that the left hemisphere inhibits the right even following damage. However, an analogous observation, again provided by experiments performed on the developing embryo, suggests an alternative explanation. Needham (1968) describes

an experiment performed by Roux in 1888 in which one cell of a two-cell frog embryo was killed by cautery. A half-embryo developed from the remaining still-living cell. This result contrasts with those of the later experiments cited by Weiss in which whole embryos developed from the isolated blastomeres of embryos at the two-cell stage. The crucial difference between the two sets of experiments may be that Roux did not isolate the cell he had killed from the living cell that remained. Clearly, the dead cell did not inhibit the living cell. Yet, by its propinquity to that cell, it constrained the living cell's course of development. The contrast between the results of the Roux experiments and those of the Driesch experiments described in the extract above is analogous to the contrast between observations of language functions in the whole brain with left-hemisphere damage, and observations of language functions in the isolated right hemisphere. Thus, it may not be the case that the damaged hemisphere inhibits the intact hemisphere. Rather, the damaged left hemisphere together with the intact right hemisphere may constitute a single system whose "equilibrium state" is incompatible with right-hemisphere control of language functions.

From the systems theoretical vantage point, the facts of cerebral dominance development are no longer paradoxical. In fact, the development of a "language center" in the right hemisphere only when it is isolated from the left is almost an expected dominance characteristic given the systems property of equifinality and given our knowledge of its operation in the developing embryo.

The dissolution of the dominance development paradox within a systems theoretical framework is more satisfactory on at least three grounds than its resolution as provided by a revision of the static dominance view. Two of the grounds have already been discussed: the inadequacy of the inhibition hypotheses and the fact that the proposed resolution invokes a construct that has been disconfirmed in the biological realm. Additionally, and perhaps most importantly, the systems description of dominance development is preferred because it is the simpler of the two descriptions. That is, it explains dominance development using principles that have already been proposed in other areas of scientific knowledge to account for properties of other systems.

The Equilibrium State Leading to Dominance Development

Adopting the systems view requires us to revise not only our account of cerebral dominance development, but also the concept of dominance itself. In this and the following section, a first approximation to a systems theoretical view of dominance will be described.

We can begin by taking stock of what has already been said about systems dynamics. The dynamics constitute the equilibrium state of the system; that is, they define the goal toward which the system tends. Being an equilibrium state, the goal exerts a regulatory influence on the current activities of the system. Again, the dynamics only establish the endpoint or the target state toward which the system works. They do not fix the route by which it will attain that state. Functions are not coded in terms of changes in particular subunits.

Therefore, although dominance typically appears to develop "in" the left hemisphere, the equilibrium state for the developing hemispheres cannot be a "left-hemisphere representation of language." The goal state must rather be one that allows for equifinality, or more generally, that allows for goal attainment despite the loss of some systems subunits--including those in the left

hemisphere. Systems theory provides a clue to the form that this developmental goal specification might take. According to Bertalanffy (1968), increasing degrees of differentiation or of hierarchical complexity is a general phylogenetic trend. One instance of the phylogenetic trend toward hierarchical complexity may be the evolution of hemispheric specialization. Differentiation is advantageous to a system because it permits more refined or reliable control of particular systems subunits. However, it also leads to a loss of flexibility. The functions that were performed by a subpart before it was damaged are not so readily compensated for if the remaining intact subunits are specialized for other functions. Therefore, according to Bertalanffy, differentiation or mechanization is never complete in a biological organism. Rather, organisms reach some compromise between the advantages of specialization and those of flexibility. Since there is no apparent reason to suppose that the human cerebral hemispheres should constitute an exception to the rule, we assume that differentiation of function in the hemispheres is likewise incomplete.

The form that the specialization/flexibility compromise often takes in biological systems is the establishment of a "leading part" (Bertalanffy, 1968). The leading part is a subunit in a system whose activities are highly influential with respect to the state of the whole system. Although each subunit in a system influences and is influenced by changes in every other subunit, they may differ in their degree of influence. A leading part is a subunit whose influence is large relative to that of other subunits. Organizers or inductors in the developing embryo are examples of leading parts.

The leading part is not considered to control the systems operations. Rather, it contributes more influentially to the character or organization of the systems dynamics than do other subunits (we recall that mechanization or differentiation is not complete). Therefore, the reactions between the leading part and other subunits are still those of mutual influence and interaction, not of cause and effect.

We might guess then that the goal state defined by the systems dynamics of language development is the establishment of a mode of cerebral function for which some subunit becomes a leading part. Under normal conditions, because of its relation to the rest of the systems subunits, some left-hemisphere subunit emerges as a leading part. If the left hemisphere is removed, a right-hemisphere subunit, with a new relation to the remaining whole, emerges as a leading part. In short, cerebral dominance provides an instance of the general rule formulated by Pattee (1970) in a discussion of biochemical structure and function: "Function is never determined by a particular structure itself, but only by the context of the organization and the environment in which this structure is embedded" (p. 119).

Dominance in the Mature Brain

In the mature brain, some hemispheric leading part acts to organize each mode of cerebral function. For instance, the left-hemisphere leading part organizes linguistic processing. However, the left-hemisphere leading part is probably not the only hemispheric leading part, nor is linguistic processing the only functional mode that a lateralized leading part acts to organize. The claim is often made that the right hemisphere is dominant for a class of tasks that, loosely stated, demand gestaltlike processing modes. If we accepted this

view, we would also assume that a right-hemisphere leading part develops ontogenetically to serve as the organizer of global or gestalt processing. The brain could then be said to engage in two broadly defined modes of processing, one organized by a left-hemisphere leading part, and one organized by a right-hemisphere leading part.

Of course, we do not want to claim that the brain has only two modes of function. The techniques used to identify these processing modes are only sensitive to lateralized modes. In systems theoretical terms, they are sensitive only to modes of function for which there is a lateralized leading part. When a subject performs a verbal task, for instance, we infer a distinct mode of processing because that mode has certain observable consequences: the subject is sensitized to visual information presented in the field contralateral to his dominant hemisphere (Kinsbourne, 1970), and to verbal information presented to his right ear (see, for example, Kimura, 1967). Furthermore, he tends to move his eyes to the right while engaging in verbal activities (Kinsbourne, 1972), and if he is speaking, he makes more movements of the right side of his body than of his left (Condon and Ogston, 1971; Kimura, 1973). Other kinds of activities, however, do not provide evidence of lateralization, or they provide evidence of a lesser degree of lateralization. These activities, we might then infer, involve processing modes for which there is no lateralized leading part, or they evoke some balance between two lateralized processing modes.

In any case, in adopting a systems view of lateralized functions, we hypothesize that such functions are whole-brain modes and that the modifier "lateralized" simply means that one hemisphere contributes more influentially than the other to the character of the processing. Since there are at least two differently lateralized processing modes, the term "dominance" can only be ascribed to a hemisphere with reference to a particular mode of functioning. Furthermore, in contrast to the static dominance view, dominance must be considered a temporary hemispheric characteristic evident only when the mode of processing with which it is associated is evoked.

Possible Tests of, and Empirical Evidence for, the Systems Hypothesis

The systems theoretical perspective on dominance in the mature brain does make testable claims to distinguish it from current views. Specifically, it claims that differentiation of function is incomplete and that the term "dominance" must therefore refer to the temporary emergence of a leading part. From this, at least two predictions follow. First, both hemispheres in the normal brain must contribute to all processing modes, even if in different degrees. Second, a nondominant hemisphere in the normal brain should contribute even to tasks it is unable to perform in isolation; a whole normal brain will therefore perform a given lateralized task more efficiently than the isolated dominant hemisphere of a callosalized individual.

If the first prediction is correct, then the right hemisphere of an intact brain must contribute to phonetic processing, even though it demonstrably cannot do so in the callosalized brain (Zaidel, 1974). Dimond's technique for assessing left- and right-hemisphere contributions to word recognition (Dimond, 1971) might be adapted to test this prediction. Dimond presents pairs of words, one each to two of the four hemiretinae, such that the words are transmitted, directly, both to the left hemisphere, both to the right hemisphere, or one to each

hemisphere. Subjects report the items more accurately, if the words are transmitted to different hemispheres than if both are transmitted to either hemisphere alone. These results suggest that the right hemisphere contributes to the processing of those words that are presented in the left visual field. If it did not so contribute, then the best condition in the experiment should be that in which both words are presented to the left, language-dominant, hemisphere.

Dimond's paradigm might be altered so that the occurrence of phonetic coding could be observed and measured. One technique for demonstrating the occurrence of phonetic coding in reading[3] involves tachistoscopic presentation of word pairs. The words are either totally unrelated or phonetically related. One of the words is marked to its left with a star, and the subject's task is to read the starred word as quickly as he can. The finding is that vocal reaction times in reading the starred word are significantly longer if members of the word pair are phonetically related than if they are unrelated. The difference in mean reaction time between the phonetically related and unrelated word pairs is a measure of phonetic interference. Merging this paradigm with Dimond's, the word pairs presented to the different hemiretinae might be phonetically related or unrelated. If the right hemisphere contributes to the phonetic processing of words transmitted directly to it, then the phonetic interference effect should be less if items are presented to different hemispheres than if both are presented to the left hemisphere. If the right hemisphere does not contribute to phonetic processing, then the interference effect should be the same in all conditions.

The second prediction has already been tested by Milner and Taylor (1972) for a mode of processing generally attributed to the right hemisphere. They showed that commissurotomized subjects are inferior to controls (with intact commissures, but comparable extracallosal brain damage) in their performance on a tactile memory task. This result obtained even when the left-hand (right-hemisphere) performances of the two groups were compared. An analogous experiment, intended to generalize the claim to left-hemisphere modes of processing, would assess the left hemisphere of a callosalized individual on some linguistic task in comparison with the whole brain of an appropriate control subject.

Finally, the systems view is compatible with, and is able to incorporate, current evidence on lateralization and on attentional effects in normal subjects. Thus, it already has a firm empirical base. However, the strongest argument for the systems view is neither its testability nor its compatibility with current evidence on dominance. It is, rather, that the view provides a perspective on dominance that effaces an anomaly in hemispheres theory--the failure of language to develop in a hemisphere with the capacity to acquire it.

REFERENCES

Asatryan, D. G. and S. G. Fel'dman. (1965) Functional tuning of the nervous system with control of movement or maintenance of a steady posture. Biophys. 10, 925-935.

---

[3] Carol Fowler and William Fischer, 1975: unpublished data.

Basser, L. (1962) Hemiplegia of early onset and the faculty of speech with special reference to the effects of early hemispherectomy. Brain 85, 427-460.

Bertalanffy, L. von. (1968) General System Theory, rev. ed. (New York: Brazillier).

Bryden, M. P. and E. Zurif. (1970) Dichotic listening in a case of agenesis of the corpus callosum. Neuropsychologia 8, 371-377.

Condon, W. S. and W. D. Ogston. (1971) Speech and body motion synchrony of the speaker-hearer. In Perception of Language, ed. by D. L. Horton and J. J. Jenkins. (Columbus, Ohio: C. E. Merrill), pp. 150-184.

Dimond, S. (1971) Hemisphere function and word registration. J. Exp. Psychol. 87, 183-186.

Entus, A. K. (1975) Hemispheric asymmetry in processing dichotically presented speech and nonspeech stimuli by infants. Paper presented at the Biennial Meeting of the Society for Research in Child Development, April, Denver, Col.

Gazzaniga, M. (1970) The Bisected Brain. (New York: Appleton-Century-Crofts).

Gazzaniga, M. and S. Hillyard. (1973) Attention mechanisms following brain bisection. In Attention and Performance IV, ed. by S. Kornblum. (New York: Academic Press), pp. 229-238.

Geffen, G., J. C. Bradshaw, and N. C. Nettleton. (1973) Attention and hemispheric differences in reaction time during simultaneous audio-visual tasks. Quart. J. Exp. Psychol. 25, 404-412.

Geschwind, N. (1969) Anatomical understanding of the aphasias. In Contributions to Clinical Neuropsychology, ed. by A. L. Benton. (Chicago: Aldine), pp. 107-128.

Gregory, R. L. (1961) The brain as an engineering problem. In Current Problems in Animal Behavior, ed. by W. H. Thorpe and O. L. Zangwill. (Cambridge, England: Cambridge University Press), pp. 307-330.

Hewitt, W. (1962) The development of the human corpus callosum. J. Anatomy 96, 355-358.

Hicks, R. (1975) Intrahemispheric response competition between verbal and unimanual performance in normal adult human males. J. Comp. Physiol. Psychol. 89, 50-60.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.

Kimura, D. (1973) Manual activity during speaking--I. Right handers. Neuropsychologia 11, 45-50.

Kinsbourne, M. (1970) The cerebral basis of lateral asymmetries in attention. Acta Psychol. 33, 193-201.

Kinsbourne, M. (1972) Eye and head turning indicates cerebral lateralization. Science 176, 539-541.

Kinsbourne, M. (1973) The control of attention by interaction between the hemispheres. In Attention and Performance IV, ed. by S. Kornblum. (New York: Academic Press), pp. 239-256.

Kinsbourne, M. (1974) Mechanisms of hemispheric interaction in man. In Hemispheric Disconnection and Cerebral Function, ed. by M. Kinsbourne and W. L. Smith. (Springfield, Ill.: Charles C Thomas), pp. 260-285.

Kinsbourne, M. and J. Cook. (1971) Generalized and lateralized effects of concurrent verbalization on a unimanual skill. Quart. J. Exp. Psychol. 23, 341-345.

Lehiste, L. (1971) Temporal compensation in a quantity language. Proceedings of the International Congress of Phonetic Sciences 7, 929-937.

Lenneberg, E. (1967) Biological Foundations of Language. (New York: John Wiley & Sons).

Milner, B., C. Branch, and T. Rasmussen. (1964) Observations on cerebral dominance. In Disorders of Language, ed. by A. V. S. deRueck and M. O'Connor. (London: Churchill), pp. 200-214.

Milner, B. and L. Taylor. (1972) Right hemisphere superiority in tactile pattern-recognition after cerebral commissurotomy: Evidence for nonverbal memory. Neuropsychologia 10, 1-15.

Molfese, D. (1972) Cerebral asymmetry in infants, children, and adults: Auditory evoked responses to speech and noise stimuli. Unpublished Ph.D. dissertation, Pennsylvania State University.

Moscovitch, M. (1973) Language and the cerebral hemispheres: Reaction time studies and their implications for models of cerebral dominance. In Communication and Affect: Language and Thought, ed. by P. Pliner, L. Krames, and T. Alloway. (New York: Academic Press), pp. 89-126.

Needham, J. (1968) Order and Life. (Cambridge, Mass.: MIT Press).

Pattee, H. (1970) The problem of biological hierarchy. In Towards a Theoretical Biology III, ed. by C. Waddington. (Chicago: Aldine), pp. 117-135.

Saul, R. E. and P. S. Scott. (1973) Compensatory mechanisms in agenesis of the corpus callosum. Neurology 23, BT68(A).

Selnes, O. (1974) The corpus callosum: Some anatomical and functional considerations with special reference to language. Brain Lang. 1, 111-139.

Smith, A. and C. Burklund. (1966) Dominant hemispherectomy: Preliminary report on neuropsychological sequelae. Science 153, 1280-1282.

Sperry, R. W. (1970) Cerebral dominance in perception. In Early Experience and Visual Information Processing in Perceptual Reading Disorders, ed. by F. A. Young and D. B. Lindsley. (Washington, D.C.: National Academy of Sciences), pp. 167-178.

Teng, E. and R. Sperry. (1973) Interhemispheric interaction during simultaneous bilateral presentation of letters and digits in commissurotomized patients. Neuropsychologia 11, 131-140.

Teng, E. and R. Sperry. (1974) Interhemispheric rivalry during simultaneous bilateral task presentation in commissurotomized patients. Cortex 10, 111-120.

Trevarthen, C. (1970) Experimental evidence for a brain-stem contribution to visual perception in man. Brain, Behavior, and Evolution 3, 338-352.

Trevarthen, C. (1974) Analysis of cerebral activities that generate and regulate consciousness in commissurotomy patients. In Hemisphere Function in the Human Brain, ed. by S. Dimond and J. G. Beaumont. (New York: Halstead Press), pp. 235-262.

Trevarthen, C. and R. Sperry. (1973) Perceptual unity of the ambient visual field in human commissurotomy patients. Brain 90, 547-570.

Weiss, P. (1969) The living system: Determinism stratified. In Beyond Reductionism, ed. by A. Koestler and J. R. Smythies. (Boston: Beacon Press), pp. 3-55.

Weiss, P. (1971) The basic concept of hierarchical systems. In Hierarchically Organized Systems in Theory and Practice, ed. by P. Weiss. (New York: Hafne), pp. 1-43.

Witelson, S. and W. Pallie. (1973) Left hemisphere specialization for language in the newborn; neuroanatomical evidence of asymmetry. Brain 96, 641-646.

Wood, C., W. Goff, and R. Day. (1971) Auditory evoked potentials during speech perception. Science 173, 1248-1251.

Zaidel, E. (1974) Language, dichotic listening and the disconnected hemispheres. Paper presented at the 15th Annual Meeting of the Psychonomic Society, Boston, November.

Zangwill, O. (1962) Dyslexia in relation to cerebral dominance. In Reading Disability, ed. by J. Money. (Baltimore: Johns Hopkins Press), pp. 103-114.

Auditory and Linguistic Processes in Speech Perception: Inferences from Six Fusions in Dichotic Listening*

James E. Cutting+

ABSTRACT

A number of phenomena in speech perception have been called fusion, but little effort has been made to compare these phenomena in a systematic fashion. The present paper examined six of them: sound localization, psychoacoustic fusion, spectral fusion, spectral/temporal fusion, phonetic feature fusion, and phonological fusion. They occur at three, perhaps four, different levels of perceptual analysis. The first two levels are characterized by perceptual integration, the other(s) by perceptual disruption and recombination. All of the fusions can be exemplified using the syllable /da/, as in dot, and all occur during dichotic listening. In each type of fusion the robustness of the fused percept is observed against variation in three parameters: the relative onset time of the two opposite-ear stimuli, their relative intensity, and their relative fundamental frequency. Patterns of results are used to confirm the arrangement of the six fusions in a hierarchy, and supporting data are summoned in an analysis of the mechanisms that underlie each with reference to speech.

Many accounts of speech and language emphasize a hierarchical structure (see, for example, Fry, 1956; Studdert-Kennedy, in press). Recently the interface between two particular levels in this system has aroused much attention: the general level logically just prior to linguistic analysis, typically called the auditory level, and the first tier of the language hierarchy logically just subsequent to that auditory analysis, the phonetic level (Wood, Goff, and Day, 1971; Studdert-Kennedy, Shankweiler, and Pisoni, 1972; Pisoni, 1973; Cutting, 1974; Wood, 1975). These and other levels of processing appear to operate in

---

*To appear in Psychological Review (1976) 83.

+Also Wesleyan University, Middletown, Conn.

Acknowledgment: This research was supported in part by a seed grant from Wesleyan University to the author. Initial portions of this research were reported by the author (Cutting, 1972, 1973) in a doctoral dissertation submitted to Yale University. I thank Ruth S. Day for her helpfulness in all stages of this enterprise; and Michael Studdert-Kennedy, Michael Turvey, Terry Halwes, Bruno Repp, and the reviewers for many suggestions to improve the paper.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

parallel, but the outcome at one level appears to be contingent on the output at a prior level (Wood, 1974, 1975; Marslen-Wilson, 1975).[1] The present paper looks at these and other levels from a new vantage point.

Information-processing analyses assume that perception takes time and that systematic disruption or complication of the process can reveal underlying properties of the system. This epistemological position often leads the researcher to paradigms of masking in both visual (Turvey, 1973) and auditory (Darwin, 1971; Massaro, 1972, 1974) modalities. Masking occurs through the rivalry of two stimuli competing for the limited processing capacities of a single processor: information is lost at a bottleneck in the perceptual system. The reciprocal process to rivalry, one equally suited to information-processing analysis, is fusion. Here information from the two stimuli is not strictly lost, but rather transformed into something new. With the outstanding exception of Julesz (1971), fusion has received little systematic attention in vision, and the phenomenon has received essentially no systematic attention in audition. The present investigation takes a small step in this direction.

One reason that little attention has been paid to auditory fusions may be that, as Julesz (1971:52) suggests, the "auditory system is basically an archaic, diffuse structure that is hard to probe." A second reason may be that seemingly too large a number of auditory phenomena have been called fusion. When one reads a given paper in this field (for example, Broadbent and Ladefoged, 1957; Sayers and Cherry, 1957; Day, 1968; Halwes, 1969; Perrott and Barry, 1969), it is clear what phenomenon is dealt with; moreover, one feels confident that the authors have properly labeled each phenomenon as a fusion. However, when one inspects the papers as a group, it is not clear that they share any common ground except for two superficial facts: they all use the word fusion in their titles and they all present their stimuli dichotically--that is, one stimulus to one ear and one to the other. Fusion is clearly not one phenomenon, but many phenomena; yet how are they related? At best, these findings appear to be just "curious binaural phenomena" (Tobias, 1972); at worst they may lead the careful reader to confusion. The purpose of this paper, then, is (a) to enumerate the different kinds of auditory fusion, (b) to arrange six of the dichotic phenomena relevant to speech processing in a hierarchy according to the processing characteristics implied by each, then (c) to confirm that arrangement by subjecting each fusion to a common set of presentational and stimulus variables that have proved important to auditory processing in general.

The list of fusions to be considered here is not intended to be exhaustive, merely organized with regard to three themes. First, all fusions here are dichotic fusions, combinations of stimuli presented to opposite ears. This stipulation eliminates temporal fusions of repeating noise patterns (Guttman and Julesz, 1963), tone patterns (van Noorden, 1975), and briefly interrupted

---

[1] More recently, however, many of the phenomena thought to characterize phonetic perception have been found to occur for music and musiclike sounds (Locke and Kellar, 1973; Bever and Chiarello, 1974; Cutting and Rosner, 1974; Blechner, Day, and Cutting, in press; Cutting, in press; Cutting, Rosner, and Foard, in press). Among other implications, these results suggest that "phoneticlike" perception is characteristic of general higher-level processing in the auditory system encompassing both speech and music.

segments of speech (Huggins, 1964, 1975). Second, all fusions reflect processes underlying the perception of speech. This eliminates consideration of centrally generated perceptions of simple pitches (Cramer and Huggins, 1958; Fourcin, 1962; Bilsen and Goldstein, 1974), patterns of pitches (Kubovy, Cutting, and McGuire, 1974), musical intervals and chords (Houtsma and Goldstein, 1972), musical illusions (Deutsch, 1975a, 1975b; Deutsch and Roll, in press), or integrated pulse trains (Huggins, 1974). Third, all fusions are exemplified by a single rule: the fused percept is different from the two dichotic inputs. This eliminates the dichotic switching-time experiments of Cherry and Taylor (1954) and Sayers and Cherry (1957), using speech stimuli, and the phenomenon of masking-level difference (see, for example, Jeffress, 1972). Masking-level difference typically eliminates itself according to the second stipulation--most often tones are imbedded in noise and interstimulus phase relations altered to yield percepts of either tone-plus-noise or noise alone. However, this is not necessarily the case. Speech stimuli can easily be imbedded in noise and their intelligibility increased through the manipulation of phase relations.

It may seem that these constraints eliminate all the possible fusions that might occur in audition. However, there are at least five others that are relevant to speech and they will be discussed with regard to a sixth and most basic type of fusion, sound localization. Since previous authors have simply called their phenomena fusion, I have taken the liberty in most cases of adding a descriptive adjective. The fusions to be considered, then, are (a) sound localization, (b) psychoacoustic fusion, (c) spectral fusion, (d) spectral/temporal fusion, (e) phonetic feature fusion, and (f) phonological fusion. For the purpose of comparability each will be discussed primarily with regard to the syllable /da/, as in dot, as shown in Figure 1. Schematic spectrograms of the dichotic stimuli are shown to suggest the manner in which items could be perceptually combined. The present paper is concerned with the pressures that can be placed on the perceptual system to inhibit fusion. All fusions are more or less subject to these pressures.

In general three variables have proved informative in the previous investigations of these fusions: relative onset time of the two stimuli, their relative intensity, and their relative fundamental frequency. Table 1 summarizes the results of that research, which used many different kinds of stimuli. Only in rare cases were the same stimuli employed to investigate more than one type of fusion or even to investigate more than one parameter within a given fusion type. The overview that follows incorporates the material from both Figure 1 and Table 1.

Sound Localization: Fusion of Two Identical Events

Sound localization has been included here as a reference point to be used when considering the other forms of fusion. All audible sounds, simple or complex, can be localized--and usually are. It is the most basic form of fusion and occurs for both speech sounds and nonspeech sounds alike. Provided that microsecond accuracy is not crucial, a convenient way to study sound localization in the laboratory is to use the same apparatus needed for studying other types of fusion: a good set of earphones, a dual-track tape recorder, and a two-channel tape with appropriate stimuli recorded on it. Approximate sound localization can be obtained using just one ear (Angell and Fite, 1901; Perrott and Elfner, 1968; Mills, 1972), but it is the two-eared phenomenon that will be discussed here.

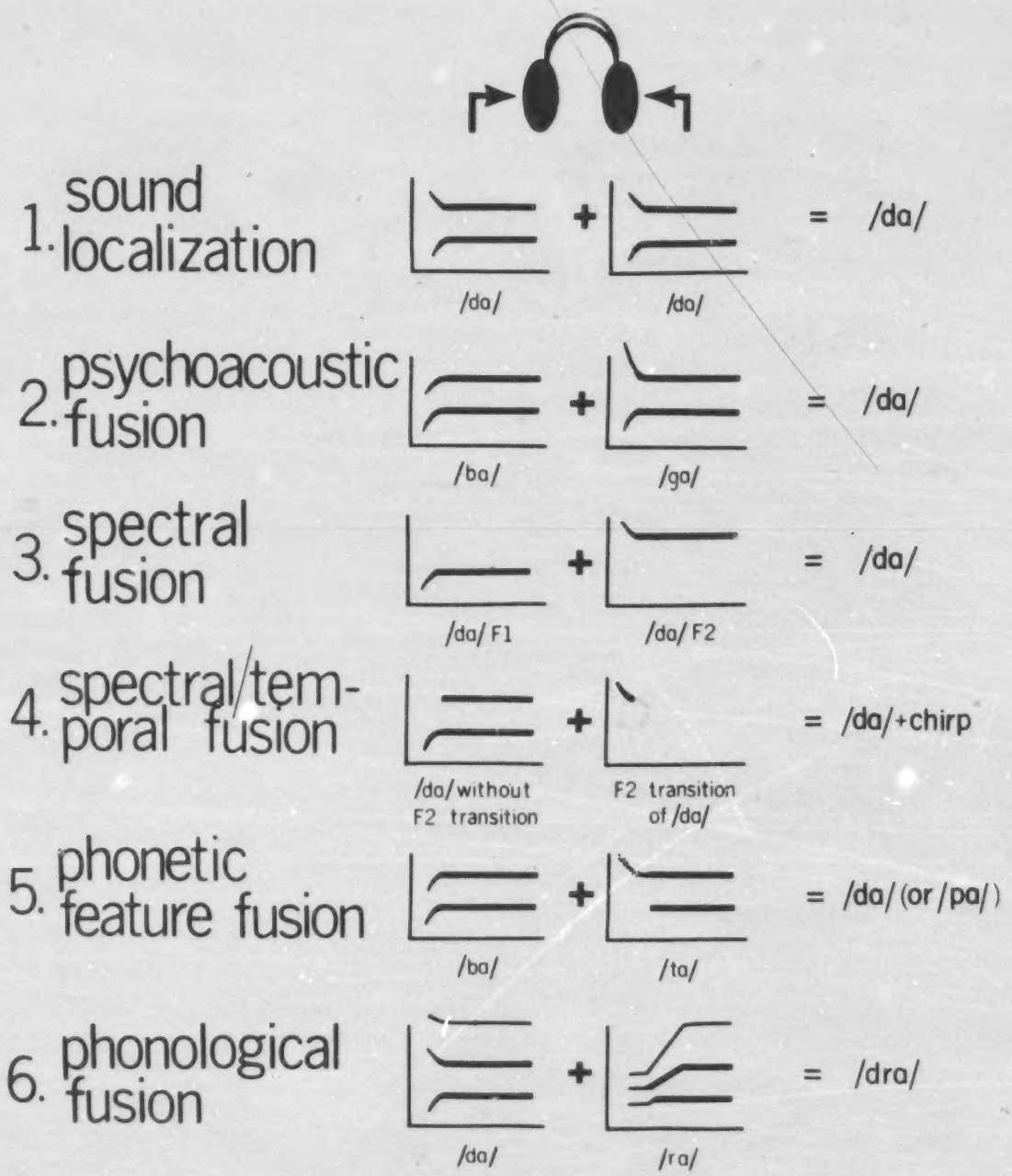


Figure 1: Schematic spectrograms of stimuli used in the six fusions (F_1 = first formant; F_2 = second formant).

TABLE 1: Upper limits of interstimulus discrepancies permitting the fusion of sounds presented to opposite ears.

Fusion type	Onset time	Intensity	Frequency
1. Sound localization	2.5 msec[a] 10 msec[d]	60 dB[b] 65 dB[e]	25 Hz[c] 80 Hz[f] <2 Hz[g]
2. Psychoacoustic fusion[h]	-	-	-
3. Spectral fusion[i,j,k]	<250 msec[i] <5 msec[n]	40 dB[l,m]	<25 Hz[j]
4. Spectral/temporal fusion[l]	-	40 dB[l]	-
5. Phonetic feature fusion[o,g]	60 msec[p] <120 msec[r]	20 dB[q]	>14 Hz[g]
6. Phonological fusion[s]	>150 msec[t] <200 msec[u]	>15 dB[u]	>20 Hz[u]

[a] Woodworth (1938:528) for clicks.

[b] Groen (1964) for binaural beats of sine waves.

[c] Licklider, Webster, and Hedlun (1950) and Perrott and Nelson (1969) for binaural beats.

[d] Cherry and Taylor (1954) for natural speech stimuli. See also Tobias (1972).

[e] Application of "cyclotean" stimuli of Kubovy et al. (1974).

[f] Perrott and Barry (1969) for sine waves near 2000 Hz; considerably greater differences possible for higher frequencies. See also Thurlow and Elfner (1969) and Tobias (1972).

[g] Halwes (1969) for synthetic speech stimuli.

[h] Gleaned from Halwes (1969).

[i] Broadbent (1955) for natural speech patterns.

[j] Broadbent and Ladefoged (1957) for synthetic speechlike patterns.

[k] Leakey, Sayers, and Cherry (1958) for nonspeech; Matzker (1959), Linden (1964), and Smith and Resnick (1972) for natural speech; Halwes (1968), Ades (1974), and Haggard (1975) for synthetic speech. Several examples cited by Tobias (1972) may also fit into this category.

[l] Rand (1974) for synthetic speech.

[m] Nye, Nearey, and Rand (1974) and Nearey and Levitt (1974) for synthetic speech.

[n] Pilot research by author using metronomelike ticks.

[o] Shankweiler and Studdert-Kennedy (1967, subsequent analysis) for synthetic speech stimuli; Studdert-Kennedy and Shankweiler (1970) for natural speech.

[p] Estimated from Studdert-Kennedy, Shankweiler, and Schulman (1970) for synthetic speech.

[q] Estimated from Cullen, Thompson, Hughes, Berlin, and Samson (1974) for natural speech stimuli.

[r] Repp (1975a, 1975b, 1975c) for synthetic speech.

[s] Day (1968) for natural speech stimuli.

[t] Day (1970a) for synthetic speech; Day and Cutting (1970) for natural speech.

[u] Cutting (1975) for synthetic speech.

The three primary parameters that affect sound localization were mentioned previously: the relative timing of the events at each ear, the relative intensity of those events, and also their relative frequency. First, consider relative timing. If one presents a brief click simultaneously to each ear, the listener reports hearing one click localized at her midline. Delaying one click by as little as 0.1 msec causes the apparent source of the percept to move away from the midline toward the ear with the leading stimulus. Delaying that click by 1 msec causes the apparent source to move to the extreme side of the auditory field away from the delayed click. With delays (onset time differences) of 2.5 msec the fused percept disintegrates and two clicks can be heard (Woodworth, 1938) shooting across the auditory field, one after the other. Apparently the effect of disintegration is postponed for longer and more complex stimuli, such as speech syllables, until relative phase differences (or onset time differences) are as great as 10 msec (Cherry and Taylor, 1954) or more (see Tobias, 1972). Thus, when two /da/s are presented to opposite ears, as much as 10 msec can separate their onsets, and a single /da/ may be heard. Experiment II is designed in part to confirm this finding.

Intensity is a second parameter affecting sound localization. Interaural differences as small as a few decibels or less easily affect the perceived locus of a sound. The problem here, however, is that unlike the potential fused percept in other fusions, the fused percept in sound localization does not "disintegrate." By making one stimulus less and less intense compared to a stimulus of constant intensity in the other ear, the locus of the percept migrates from the midline toward the ear of the most intensive stimulus. Most importantly the difference between percepts in some binaural and monaural conditions is negligible, if detectable at all. I have found that when Stimulus A is presented to one ear at a comfortable listening level and to the other ear at 20 dB lower, and when Stimulus A is presented to that ear at that comfortable listening level and no stimulus is presented to the other ear, one cannot consistently hear the difference between the two trials. To get around this problem, at least with regard to sine waves, one can look to the phenomenon of binaural beats: a "whooshing" or "roughness" percept, depending on the frequency difference between sine waves (to be discussed below). This phenomenon is perceived through the cross-correlation of opposite-ear signals. Groen (1964) presents data that suggest the intensity of the signals can differ by as much as 60 dB and a "pulling" can still be heard. Although Groen's procedure is not clear, I have replicated that result using the stimuli of Kubovy et al. (1974). In this case a melody is heard through the cross correlation and subsequent localization of the melodic elements as spatially distinct from a background of complex noise. No melody can be heard in a single stimulus because physically it is not present (which distinguishes the phenomenon from masking-level difference). The percept is robust enough so that if one stimulus is presented to one ear at 100 dB re 20 $\mu N/m^2$ and the other stimulus to the other ear at 35 dB, a faint melody can still be heard and identified. The fact that it can be heard suggests that the sound localization process is still functional at interaural intensity differences as great as 65 dB. Nevertheless, intensity is not considered relevant to the fused percept when /da/ is presented to both ears, because the percept never ceases to be /da/.

The third parameter is frequency. Sine waves may differ in frequency by as much as 25 Hz in certain frequency ranges, and a fused percept, the "roughness" of binaural beats, can be maintained. The principle here is that stimuli differing slightly in frequency can be thought of as stimuli with identical

frequencies but with constantly changing phase relations. Differences of 3 Hz or less can be heard as an oscillating stimulus whirling around the head in an elliptical path (Oster, 1973). Greater frequency differences are heard as roughness until the two tones break apart. Outside the realm of binaural beats, Perrott and Barry (1969) found that dichotic tones of considerably greater frequency differences can be heard as a single stimulus, especially above about 2000 Hz. However, when the signals are complex and periodic the pitch is typically much lower--for speech sounds, in particular, a fundamental frequency of 100 Hz is not uncommon in male speakers. It can be easily demonstrated that a /da/ at 100 Hz presented to one ear and a /da/ of 102 Hz presented to the other are heard as two events. Experiment IV is designed in part to confirm this finding.

Psychoacoustic Fusion: Fusion of Proximal Acoustic Features by Perceptual Averaging

Unlike sound localization, about which many volumes have been written, little has been written about psychoacoustic fusion. Here acoustic features from opposite-ear stimuli appear to be averaged to yield an intermediate result. The phenomenon could logically occur for many different kinds of sounds, both speech and nonspeech. Its existence is gleaned from my own experimentation with dichotic synthetic speech stimuli and from a large table of data presented by Halwes (1969:61). Perhaps, the best way to demonstrate the phenomenon is to consider the stimuli in Figure 1. If this particular /ba/ is presented to one ear and this particular /ga/ to the other, the listener often reports hearing a single item, /da/. Notice that these stimuli differ only in the direction and extent of the transition in the second formant, the resonance of highest frequency for these items (Delattre, Liberman, and Cooper, 1955). Note further that the second-formant transitions are arrayed such that /da/ lies between /ba/ and /ga/, and that an "average" of the two extreme items would fall very close to /da/. To date little is known about this type of fusion, in part because it is less likely to occur with the more complex and more widely differing natural speech syllables. Therefore, Experiment I is designed to demonstrate the psychoacoustic nature of the phenomenon. There the relation between this fusion and a similar phenomenon known as the "feature sharing effect" will also be considered (see Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy et al., 1972; Pisoni and McNabb, 1974; Repp, 1975a). Experiments II, III, and IV are designed to observe the effect of the three crucial variables on psychoacoustic fusion.

Spectral Fusion: Fusion of Different Spectral Parts of the Same Signal

Broadbent (1955) and Broadbent and Ladefoged (1957) reported this third phenomenon, which I call spectral fusion. It occurs when different spectral ranges of the same signal are presented to opposite ears. A given stimulus, for example, is filtered into two parts: one containing only the low frequencies and the other containing only the high frequencies. Each is presented separately but simultaneously to opposite ears. The listener invariably reports hearing the original stimulus, as if it had undergone no special treatment. In his initial study, Broadbent found that this fusion readily occurs for complex stimuli of many types, nonspeech sounds, such as metronome ticks, and speech sounds as well. Moreover, when listeners were informed about the nature of the stimuli and asked to report which ear had the low-frequency sounds and which ear had the high frequencies, they performed at chance level.

Relative timing is an important parameter in spectral fusion. Broadbent (1955) found that arrival time differences of 250 msec were sufficient to disrupt the fused percept. My own pilot research suggests that this interval may be at least an order of magnitude too large. For example, when the different spectral portions of metronomelike ticks are offset by as little as 5 msec, the listener hears two sets of ticks, not one. As in sound localization, the temporal differences tolerable in spectral fusion may be greater for more complex sounds such as speech items. Therefore, Experiment II is directed at finding the relative onset time limits allowable when a speech syllable /da/ is spectrally split and its first formant presented to one ear and the second formant to the other. For generality, the syllables /ba/ and /ga/ will also be used.

The effect of relative intensity in spectral fusion has been explored systematically by Rand (1974; see also Nearey and Levitt, 1974; Nye, Nearey, and Rand, 1974). The most interesting case occurs when the second (and higher) formants, presented to one ear, are decreased in amplitude with respect to the first formant, presented to the other ear. Rand found that decreases of as much as 40 dB have little effect on identifiability of /ba, da, ga/. This large effect is particularly surprising since attenuations in the upper formants of only 30 dB are sufficient to render the syllables unrecognizable when all formants are presented to both ears. Rand termed the phenomenon "dichotic release from masking" and the release is clearly substantial. The emphasis in the present paper is not on masking but on fusion, but since Rand used stimuli very similar to those used in the present studies and since Nye et al. (1974) have already replicated those results, intensity effects will not be explored further.

Fundamental frequency is also an important parameter in spectral fusion. Broadbent and Ladefoged (1957) found that the fundamental frequencies of the to-be-fused stimuli must be identical for fusion to occur (that is, for one item to be heard). Differences of 25 Hz inhibited fusion, and Halwes (1969) suggests that differences of 2 Hz may inhibit fusion as well. But there appear to be two types of fusion here: one concerns the number of items heard, one or two, and the other concerns the identity of the stimulus. While the effect of differences in pitch between the two component stimuli on the identity of the fused percept are considered in Experiment IV, the effect of fundamental frequency on the numerosity of percepts is considered in Experiment V.

Spectral/Temporal Fusion: Perceptual Construction of Phonemes from Speech and Nonspeech Stimuli

Rand (1974) discovered a fourth type of fusion. In addition to dividing the stimulus spectrally and presenting those portions to either ear, as noted previously, he divided the speech syllable both spectrally and temporally. Two-formant renditions of his stimuli are shown schematically in Figure 1. One stimulus is simply the second-formant transition excised from the syllable and presented in isolation. Mattingly, Liberman, Syrdal, and Halwes (1971) noted that these brief glissandi sound like the discrete elements of birdsong, and they dubbed them "chirps." The second stimulus is the remainder of the syllable without the second-formant transition. It should be noted that the transitionless /da/--that is, the speech sound without a second-formant transition--is not identifiable as /da/ when presented in isolation: instead it is almost 85 percent identifiable as /ba/. This appears to result from the upward spread of harmonics of the first-formant transition, which may mimic the now-absent second-formant transition in such manner as to cue /b/.

Perhaps the most interesting aspect of spectral/temporal fusion, and the aspect that distinguishes it from the logically similar spectral fusion, is that the listener hears more than one auditory event. He does not hear two speech sounds. Instead, he hears one speech sound, /da/, and one nonspeech sound, a chirp. Note that the perceptual whole is greater than the sum of the parts: the listener "hears" the second-formant transition in two different forms at the same time. One form is in the complete syllable /da/, which would sound more like /ba/ without it. The second form is similar to the transition heard in isolation--a nonspeech chirp. Thus, spectral/temporal fusion is more complex phenomenologically than the three fusions previously considered. It may be possible for it to occur for nonspeech sounds (perhaps a complex nonspeech sound could be segmented spectrally and temporally in the same manner and analogous percepts obtained). Nevertheless, it will be discussed here exclusively with respect to speech.

Of the three relevant parameters--onset time, intensity, and frequency--only intensity has been explored thus far. As in spectral fusion, Rand (1974) attenuated the isolated second-formant transitions of /ba, da, ga/ by as much as 40 dB and identifiability was largely unimpaired. This result was in marked contrast to the condition in which the syllable remained as an integral whole, but with the second-formant transition attenuated as before; 30 dB was sufficient to impair identification. As in spectral fusion, the intensity data are not replicated here, but the effects of differences in relative onset time and frequency are explored in Experiments II and IV, respectively. Again, /da/ will be used as a reference syllable, but /ba/ and /ga/ will also be used for the sake of generality.

Phonetic Feature Fusion: Recombination of Phonetic Feature Values by Perceptual Misassignment

With this fifth type of fusion we move to a domain that exclusively belongs to speech. Halwes (1969), Studdert-Kennedy and Shankweiler (1970), and Repp (1975a) have reported that misassignment to phonetic feature values often occurs in the dichotic competition of certain stop-vowel syllables. This "blending" can be thought of as phonetic feature fusion. Figure 1 shows that when /ba/ is presented to one ear and /ta/ to the other, the listener often reports hearing a syllable not presented. The most frequent errors are the blends /da/ and /pa/. Here the listener combines the voicing feature value of one stimulus with the place feature value of the other. For example, the voicing value of /b/ is combined with the place value of /t/ and the result is the fusion /d/.

Consider a stimulus repertory of six items: /ba, da, ga, pa, ta, ka/. On a particular trial when /ba/ and /ta/ are presented to opposite ears, and when the subject is asked to report what he or she hears, three types of responses can occur: correct responses /ba/ or /ta/, blend responses /da/ or /pa/, and anomalous responses /ga/ or /ka/. The last two items are anomalous because, although they share the voicing value with one item in the stimulus pair, neither shares place values. Using natural speech stimuli, Studdert-Kennedy and Shankweiler (1970) found that the ratio of blends (phonetic feature fusions) to anomalous responses was about 2:1, a rate significantly greater than chance. Halwes (1969:65) found that synthetic speech items occur at a rate of 10:1, or better (p. 64). Studdert-Kennedy et al. (1972) found these fusions to occur even when the vowels of the two stimuli differed markedly.

Evidence for the effect of relative onset on phonetic feature fusion is only indirect. Studdert-Kennedy et al. (1970) found that errors in identification occur more often when competing pairs of stimuli are slightly time-staggered with respect to their relative onset than when they are simultaneous. The effect decreases substantially for relative onset times of greater than 70 msec or so. The maximum error rate occurs for asynchronies of about 40 msec. If we assume that the ratio of blend responses to anomalous responses is constant for different leads, maximum phonetic feature fusions should occur at about 40 msec lead time, but should fall off rather rapidly thereafter. Experiment II was designed in part to confirm these predictions.

Evidence for the effect of relative intensity on phonetic feature fusion is equally indirect. The data of Cullen et al. (1974) demonstrate that dichotic items can compete with one another, that is, yield substantial error rates, when the two items differ by as much as 20 dB. If we assume that the ratio of blend responses to anomalous responses is constant for different intensities, phonetic feature fusions should continue to occur rather readily until intensity differences between the two stimuli are greater than 20 dB, at which point errors largely cease. Errors (and fusions) should be greatest when the two stimuli have the same intensity. Experiment III was designed in part to confirm these predictions.

The effect of fundamental frequency differences between the stimuli is better known for phonetic feature fusion. Halwes (1969) found these fusions to occur almost as frequently when the two competing stimuli had different fundamental frequency as when they had the same fundamental. Experiment IV extends the frequency differences well beyond the 14 Hz of Halwes to observe the effect on reported fusions.

Phonological Fusion: Perceptual Construction of Phoneme Clusters

Phonological fusion occurs when two imputs, each of n phonemes, yield a response of n + 1 phonemes. Day (1968) found that compatible phoneme strings, one beginning with a stop and the other a liquid, could be fused into one unit: given PAHDUCT and RAHDUCT presented to opposite ears, the subject often hears PRODUCT. One of the unique aspects of phonological fusion is that, unlike psychoacoustic fusion and phonetic feature fusion, two stimuli that contain different phonetic segments are presented at the same time, and yet they are combined to form a new percept that is longer and linguistically more complex than either of the two inputs. Another unique aspect of this fusion is that the order in which the phonemes fuse is phonologically ruled: BANKET/LANKET yields BLANKET, not LBANKET. Note that in English, initial stop + liquid clusters occur frequently but that initial liquid + stop clusters never occur: /b, d, g, p, t, k/ can typically precede /l, r/, but the reverse is never true at the beginning of a syllable. When these phonological constraints are lifted, fusion can occur in both directions: thus, TASS/TACK can yield both TACKS and TASK responses (Day, 1970b). Other linguistic influences on phonological fusion are discussed by Cutting (1975) and Cutting and Day (1975).

The effects of relative onset time have been explored by Day (1970b), Day and Cutting (1970), and Cutting (1975). Their results show that phonological fusion is remarkably tolerant of differences in onset time. When no lead times are greater than 150 msec, fusion occurs readily at all leads. When much longer

lead times are used, fusion remains frequent at all relative onsets of 100 msec and less. Factors such as whether the to-be-fused stimuli are natural or synthetic speech, whether the inputs are words or nonwords, and whether the stimuli are monosyllabic or disyllabic appear to play a role. Experiment II explores the fusion of synthetic speech items /da/-/ra/ and /ba/-/la/, varied in relative onsets similar to those of other fusions. Cutting (1975) found that phonological fusion did not decrease with intensity and frequency differences between fusible stimuli of as much as 15 dB and 20 Hz. Experiments III and IV extend the ranges of those differences to explore possible effects on the fused percept.

PURPOSE OF THE PRESENT EXPERIMENTS

The purpose of the studies that follow is fivefold. First, Experiment I is designed to demonstrate that psychoacoustic fusion is a separate phenomenon resulting from the perceptual averaging of acoustic features. Second, Experiments II through IV are designed to replicate the results found by previous studies (Table 1) using, as much as possible, the same stimulus or percept for each (/da/), and using the same group of listeners. Third, those experiments are designed to fill in the data gaps, particularly with regard to confirming estimates for phonetic feature fusion. Fourth, Experiment V and the discussion that follows it are directed at the interactions of different fusions. And fifth, from the results of all studies the different fusions will be considered with respect to the types of mechanisms that must exist at different processing levels and to their relevance in speech perception.

EXPERIMENT I: DEMONSTRATION OF PSYCHOACOUSTIC FUSION

Of the six fusions, least is known about psychoacoustic fusion. Its existence is gleaned from a single table presented by Halwes (1969) and he does not discuss this particular phenomenon. Although several types of different synthetic syllable pairs can yield a single percept ambiguous between the two dichotic items--/ba/-/ma/, /ra/-/la/, /pa/-/ba/, to name a few--it may be only the pair /ba/-/ga/ that will frequently yield a percept, /da/, different from either of the two inputs. What causes such fusion responses? Two hypotheses appear tenable. First, and supporting the notion that this fusion is psychoacoustic, the listener may hear /da/ simply because the acoustic average of the second-formant transitions for /ba/ and /ga/ happen to fall in the middle of the /da/ range. A second, alternative view is that the perceptual averaging may be more abstract. Perhaps linguistic information is extracted from the dichotic syllables with respect to place of articulation (see Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967): /b/ is labial, /g/ is velar, and the articulatory mean between the two gestures is close to the alveolar /d/. These two hypotheses have different predictions about what happens to the fused percept when acoustic variation takes place within the /ba/ and /ga/ inputs. The first hypothesis predicts that the percentage of /da/ responses would vary according to the acoustic structure of the inputs; the second hypothesis, on the other hand, predicts no change in the number of /da/ responses since all inputs are good exemplars of /ba/ and /ga/, and /da/ is always an articulatory mean between the two.

Method

Four stimuli were generated on the Haskins Laboratories parallel resonance synthesizer. Two were /ba/ and two were /ga/. All four were two-formant,

300-msec stimuli with a constant fundamental frequency of 100 Hz. First formants were centered at 740 Hz and second formants at 1620 Hz. First-formant transitions were 50 msec in duration, rising in frequency, and identical for all four items. Second-formant transitions were 70 msec in duration and varied in slope and direction, as shown in Figure 2.[2] The two stimuli nearest the /da/ boundaries are called $\underline{b}^1$ and $\underline{g}^1$ (for /ba/ and /ga/, respectively), whereas the two stimuli farthest from the boundaries are $\underline{b}^2$ and $\underline{g}^2$. Start frequencies for the second-formant transitions were 1232 and 1386 Hz for the two /ba/ stimuli, and 1996 and 2156 Hz for the two /ga/ stimuli. Boundaries and start frequencies are based on the findings of Mattingly et al. (1971), who used very similar stimuli. Pretesting determined that all items were at least 90 percent identifiable as the appropriate /b/ or /g/. Items were digitized and stored on disk file for the preparation of dichotic tapes (Cooper and Mattingly, 1969).

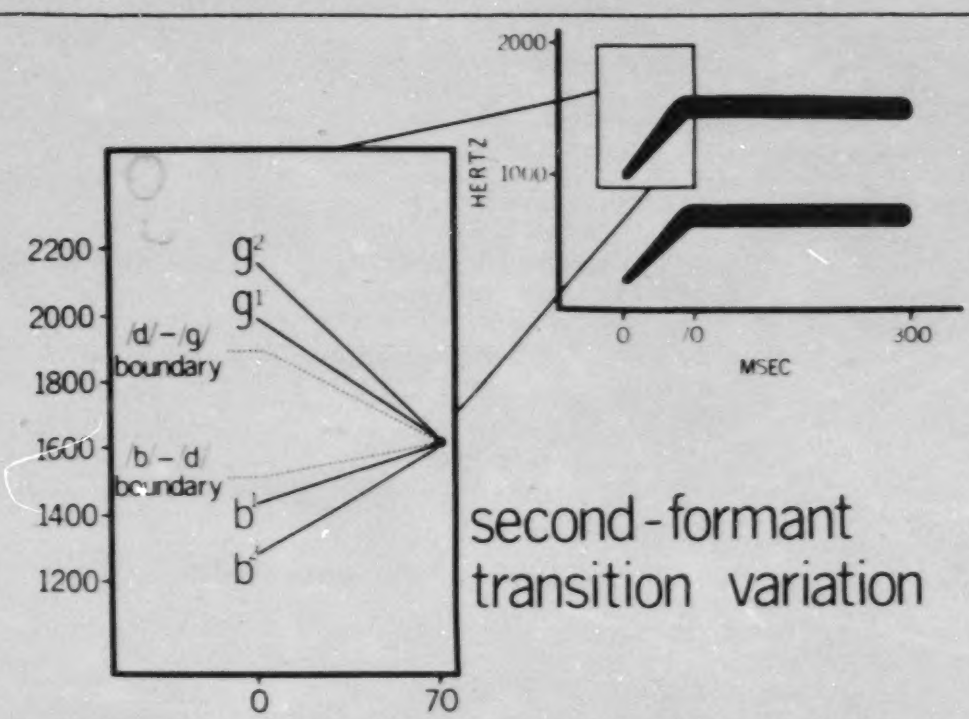


Figure 2: Schematic spectrogram and representations of four stimuli, two /ba/s and two /da/s, used in Experiment I for psychoacoustic fusion.

Four dichotic pairs were assembled: $\underline{b}^1$-$\underline{g}^1$, $\underline{b}^2$-$\underline{g}^2$, $\underline{b}^1$-$\underline{g}^2$, and $\underline{b}^2$-$\underline{g}^1$. Each of these pairs was repeated ten times in a random sequence, with 3 sec between pairs and with channel assignments properly counterbalanced. Stimuli were reconverted into analog form at the time of recording the test tape. Ten Wesleyan University undergraduates participated in the task as part of a course project;

[2] These transitions are longer than those typically found in synthetic speech syllables, but results of preliminary tests suggested that longer transitions facilitate psychoacoustic fusion. For other effects with longer transitions, see Tallal and Piercy (1974, 1975), whose data support the notion that transition duration has auditory consequences independent of phonemic consequences.

four others had been eliminated because they failed to identify the stimuli as desired. Each was a native American English speaker with no history of hearing difficulty, no experience at dichotic listening, and limited experience with synthetic speech. The tape was played on a Crown CX-822 tape recorder, and signals sent through a set of attenuators to matched Telephonics headphones (Model TDH39). Stimuli were presented at approximately 80 dB re 20 $\mu N/m^2$. Earphone-to-ear assignments were counterbalanced across listeners. They wrote down B, D, or G to identify the item that they heard, and remained uninformed that two items were actually presented. Note that no /da/ items were presented. Except when noted otherwise, procedure and apparatus were the same for all studies.

Results and Discussion

The percent /ba/, /da/, and /ga/ responses for the four types of dichotic pairs is shown in Table 2. The largest number of /da/ fusions occurred for the $\underline{b}^1$-$\underline{g}^1$ pair, the fewest for the $\underline{b}^2$-$\underline{g}^2$ pair, with intermediate fusion scores for the other two pairs. Of paramount interest is the fact that not only did /da/ fusions decrease for the dichotic pair whose second-formant transitions were farthest from the /d/ boundaries, but both /ba/ and /ga/ responses increased as well. The difference between the number of /da/ responses for $\underline{b}^1$-$\underline{g}^1$ and $\underline{b}^2$-$\underline{g}^2$ pairs is significant by Wilcoxon matched-pairs signed-ranks test [$\underline{T}(8) = 1.5$, $\underline{p} < .02$, two-tailed], and demonstrates that the proximity of the second-formant transitions in the to-be-fused pair to the /da/ boundaries is important to the phenomenon. Thus, simple averaging of formant transitions is insufficient to explain consistent psychoacoustic fusions; it is the averaging of optimally similar transitions that is crucial. This result also suggests a close relationship between rivalry and fusion, as noted earlier.[3] The responses for the other two dichotic pairs are predictable from the first results. The $\underline{b}^1$-$\underline{g}^2$ pair has an intermediate number of /da/ fusions and a smaller number of /ba/ responses. The $\underline{b}^2$-$\underline{g}^1$ pair also has an intermediate number of /da/ fusions but has an increase in the /ba/ responses, largely at the cost of /ga/. Subtracting the number of /ga/ responses from the number of /ba/ responses for each subject, this shift pattern is statistically robust [$\underline{T}(7) = 0$, $\underline{p} < .02$]. It would appear that this fusion is psychoacoustic rather than psycholinguistic since it is quite sensitive to phonemically irrelevant acoustic variation in the stimuli.[4]

---

[3] Fusion and rivalry are clearly not exclusive alternatives but interact in a probabilistic fashion. In Experiment I, as in all studies presented here, I have tried to maximize the probability of fusion in most conditions. The terms rivalry and fusion as used here are intended to parallel their use by Julesz (1971:23). Whether there is a suppression of rivalry by various fusion mechanisms in audition like that proposed in vision (Kaufman, 1963, 1974; Julesz, 1971:218-220) is not known, and is beyond the scope of the present investigation. It may be that fusion and rivalry can occur simultaneously in audition or in vision.

[4] In a study done independently at the time the present investigation was being readied for publication, Repp (1975d) found essentially the same results as those reported here, using /bae/-/gae/ dichotic pairs. From the results of several experiments on psychoacoustic fusion, he reaches many of the same, but some different, conclusions that I reach.

TABLE 2: Percent /ba/, /da/, and /ga/ responses for four pairings of dichotic stimuli in Experiment I (psychoacoustic fusion).

Pair	Response		
	B	D	G
$\underline{b}^1$-$\underline{g}^1$	9	56	35
$\underline{b}^2$-$\underline{g}^2$	15	24	61
$\underline{b}^1$-$\underline{g}^2$	5	32	63
$\underline{b}^2$-$\underline{g}^1$	21	47	32

A phenomenon similar to psychoacoustic fusion in certain respects is the "feature sharing effect" as formulated by Studdert-Kennedy et al. (1972) and explored parametrically by Pisoni and McNabb (1974). The emphasis of Pisoni and McNabb, like that of Rand (1974) for another type of fusion, is on masking; here, of course, the emphasis is on fusion. They found that given a target syllable, such as /ba/, presented to one ear and a mask, such as /ga/, presented to the other ear, listeners made few errors identifying the target regardless of the interval between the onsets of the target and mask. They found that for /b/-/g/ pairs it made no difference whether the target and mask shared the same vowel or had different vowels, but they did note that "the more similar the vowels of the two syllables, the more likely they are to 'fuse' or integrate into one perceptual unit so that the listener had difficulty assigning the correct auditory features to the appropriate stimulus" (p. 357). The fusion that they allude to is most likely the psychoacoustic fusion reported here. Given a /ba/-/ga/ target-mask pair, however, they found few /da/ responses to occur even when the items had simultaneous onsets. The reason for this result most likely stems from subject expectations and the difference between the two procedures. For example, their listeners knew that targets would be presented to one ear, masks to the other, that two items would be presented on every trial, and that they were to report the identity of the first item; here, on the other hand, there were no targets or masks, listeners did not know that two items were presented on each trial, and they were simply to report what they heard. In the present study, relative onset time was not varied as in the Pisoni and McNabb study; however, Experiment II varies onset time in a similar fashion. Just as I have concluded that psychoacoustic fusion occurs prior to phonetic processing, Pisoni and McNabb (1974) conclude that the feature sharing effect is also prephonetic.

The present experiment confirms the existence of psychoacoustic fusion and also strongly suggests that its nature is not linguistic but rather a perceptual integration of acoustic features. The major thrust of the paper, however, stems from those experiments that follow--replication and exploration of the effects of varying relative onset time, relative intensity, and relative frequency on the six fusions outlined previously.

EXPERIMENTS II-V: GENERAL METHODOLOGY

Overview

Sixteen different brief experiments were conducted to study the six fusions. All dealt with the syllable /da/, either as a stimulus or as a potential percept, as shown schematically in Figure 1. In general, three experiments were directed at each type of fusion: in one, the relative onset time of the dichotic stimuli was varied; in a second, the relative intensity was varied; and in a third, the relative fundamental frequency was varied. For simplicity's sake, rather than numbering each separate demonstration as an experiment, all those dealing with relative onset time are considered as part of Experiment II, those dealing with relative intensity as Experiment III, and those dealing with relative frequency as Experiment IV. Experiment V deals with interaction of different fusions.

The same ten listeners participated in these experiments as participated in Experiment I. Because of the great number of separate studies, counterbalancing of test order was not attempted: instead, all subjects listened first to the three tests pertaining to phonological fusion, then those to sound localization, psychoacoustic fusion, spectral fusion, spectral/temporal fusion, and phonetic feature fusion, respectively.

Stimuli

Six speech syllables--/ba, da, ga, ta, la, ra/--were generated in several renditions using the Haskins Laboratories parallel resonance synthesizer. The /ba/ and /ga/ stimuli were the $\underline{b}^1$ and $\underline{g}^1$ items used in Experiment I. All stimuli were 300 msec in duration and shared the same /a/ vowel used previously. The stop-consonant stimuli (those beginning with /b/, /d/, /g/, and /t/) consisted of three formants. For the stop stimuli first- and second-formant transitions were 50 and 70 msec in duration, respectively. Start frequency of the second-formant transitions for /da/ and /ta/ was 1695 Hz. All voiced stops (/b/, /d/, and /g/) had a voice-onset-time (VOT) value of 0 msec, while the voiceless stop (/t/) was aspirated with a VOT of +70 msec (see Lisker and Abramson, 1964). Liquid items began with 50 msec of steady-state resonance in all formants, followed by 100 msec of transitions in the second and third formants (only 20 msec in the first formant), followed by the vowel resonances. An open-response pretest showed that each item was identified correctly on at least 86 percent of the trials.

The standard forms of all these items had a pitch of 100 Hz (like that of an adult male), and an intensity of approximately 80 dB re 20 $\mu N/m^2$. Nonstandard forms were generated with frequencies of 102, 120, and 180 Hz, and with intensities ranging downward to 40 dB in 5-dB steps. For spectral fusion and for spectral/temporal fusion, /ba/, /da/, and /ga/ were also parsed into separate parts, as shown in Figure 1: in spectral fusion items were generated as separate formants, and in spectral/temporal fusion the second-formant tarnsition was isolated from the remainder of the syllable. All stimuli were digitized and stored on computer disk file for the preparation of dichotic tapes. Except as noted below, procedures were identical to those of Experiment I.

EXPERIMENT II: RELATIVE ONSET TIME IN SIX FUSIONS

Method

Six different randomly ordered sequences of dichotic pairs were recorded, one for each type of dichotic fusion. Relative onset times were chosen with regard to the temporal range for which pretesting had determined each fusion most sensitive.

1. Sound localization. Tokens of the standard form of the stimulus /da/ (100 Hz and 80 dB) were recorded on both channels of audio tape. The items could have synchronous onsets (0-msec lead time) or asynchronous onsets. Ten asynchronous onsets were selected: 1, 2, 3, 4, 5, 10, 20, 40, 80, and 160 msec lead times. A sequence of 24 items was recorded: (10 asynchronous leads) × (2 lead time configurations, Channel A leading Channel B and vice versa) + (4 simultaneous-onset pairs). Listeners were told to write down how many items they heard--one or two--paying no special regard to the identity of the stimuli heard.

2. Psychoacoustic fusion. Tokens of the standard /ba/ were recorded on one channel and tokens of the standard /ga/ on the other. Nine lead times were selected: 0, 1, 2, 5, 10, 20, 40, 80, and 160 msec. A sequence of 36 dichotic pairs was recorded: (9 leads) × (2 lead time configurations) × (2 channel assignments, /ba/ to Channel A and /ga/ to Channel B, and vice versa). Listeners were instructed to write down the initial consonant of the syllable that they heard most clearly, choosing from among the voiced stops B, D, or G. Note that no /da/ stimuli were actually presented.

3. Spectral fusion. The first formants of the standard items /ba/, /da/, and /ga/ were recorded on one channel and the second formants on the other. Six lead times were selected: 0, 10, 20, 40, 80, and 160 msec. A sequence of 72 items was recorded: (3 stimulus pairs, first and second formants for /ba/, /da/, and /ga/) × (6 lead times) × (2 lead time configurations) × (2 channel assignments). Listeners wrote down the initial consonant that they heard most clearly, B, D, or G.

4. Spectral/temporal fusion. Each of the 70-msec second-formant transitions was excised from the three standard syllables, /ba/, /da/, and /ga/, and recorded on one channel, and the remainder of the syllables recorded on the other. A sequence of 72 items was recorded following the same format as the spectral fusion sequence. Again, listeners wrote down B, D, or G.

5. Phonetic feature fusion. The standard form of /ba/ was recorded on one channel, and the standard form of /ta/ on the other. Seven lead times were selected: 0, 5, 10, 20, 40, 80, and 160 msec. A sequence of 84 items was recorded: (7 leads) × (2 lead time configurations) × (2 channel arrangements) × (3 observations per pair). Listeners wrote down the initial consonant of the syllable that they heard most clearly, choosing from among B, D, P, and T. Note that no /da/ or /pa/ stimuli were actually presented.

6. Phonological fusion. Two types of dichotic pairs were recorded on opposite channels: /ba/ and /la/, and /da/ and /ra/, all of standard form. Five leads were selected: 0, 20, 40, 80, and 160 msec. A sequence of 40 items

was recorded: (2 fusible dichotic pairs) × (5 leads) × (2 lead time configurations) × (2 channel assignments). Listeners were instructed to write down whatever they heard, following Day (1968).

Results and Discussion

Relative onset time is a crucial variable for all six fusions, as shown in Figure 3. In general the greater the interval between onsets of members of the dichotic pair, the less frequently fusion occurs; or, inversely, the greater the probability of the perceptual disintegration of the fused percept. All results will be discussed with regard to that relative onset time at which fusions first occur significantly less frequently than that for simultaneous onset pairs, as measured by a Wilcoxon matched-pair signed-ranks test ($p < .05$).

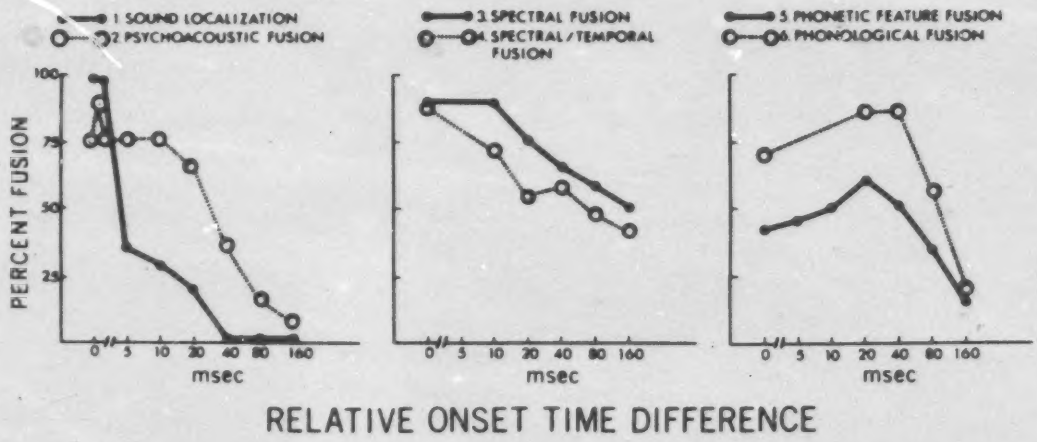


Figure 3: The effect of relative onset time in six fusions. Data from Experiment II.

For sound localization, the fusions, or one-item responses, decreased precipitously with relative onset times as small as 4 and 5 msec. (In the left-hand panel of the figure, onset differences of 2 and 3 msec are combined, as well as those of 4 and 5 msec, because there was little difference between them.) This estimate is slightly smaller than that given by Cherry and Taylor (1954), but it should be noted that some subjects continued to report that they heard only one item for onset asynchronies as great as 20 msec. No one-item responses were given for relative onsets of 40 msec or greater. In the same panel, the data for psychoacoustic fusion present a different pattern. Fusions, D responses given a

/ba/-/ga/ dichotic pair, occurred readily for all trials with relative onsets as great as 20 msec, but dropped considerably for those with greater asynchronies. There was no significant effect of the order in which the two stimuli arrived: /ba/-leading-/ga/ and /ga/-leading-/ba/ trials yielded equally frequent /da/ responses.

The pattern of fused responses for spectral fusion and spectral/temporal fusion (B, D, or G for the components of /ba, da, ga/, respectively) were markedly parallel. They are shown in the central panel of the same figure. Significant decreases in the probabilities of fusion responses occur by 40 msec in spectral fusion and by 20 msec in spectral/temporal fusion. In both, the results presented are those summed over /ba/, /da/, and /ga/ items. Combined, these functions asymptote at slightly greater than 33 percent performance. Listeners choose from among three responses, one of which must be correct (unlike psychoacoustic fusion where there were no /da/ items presented). Moreover, the first formant alone sounds somewhat like /ba/ itself: fully 85 percent of all responses for both fusions at 160-msec asynchronies were /ba/. This fact contributes to the relatively high asymptote of the combined functions. There was no significant effect of the order of arrival of the dichotic items for either fusion, but B and D fusions were more frequent than G fusions. This result will be considered in the general discussion.

Phonetic feature fusion and phonological fusion present patterns of fused responses not seen for the others. In both, fusions are slightly more frequent at brief relative onset times than at simultaneous onsets or at longer relative onsets. For phonetic feature fusion this effect is significant: the 20-msec asynchrony yields more fused responses than those of 0, 80, and 160 msec. For phonological fusion, however, the increase is not significant. In phonetic feature fusion, fused responses were equally frequent regardless of whether /ba/ began before /ta/, or vice versa. For all asynchronies, /pa/ fusions were more frequent than /da/ fusions: the overall ratio was 4:1. In phonological fusion the nonlinear pattern of fusion responses is slightly complicated by the fact that when, for example, /da/ began before /ra/ fusions were more frequent than when /ra/ began before /da/. Whereas the effect was not significant here, the trend is similar to that found by Cutting (1975, Experiment III; but see Day, 1970b; Day and Cutting, 1970). Fused responses were more frequent for /ba/-/la/ pairs than for /da/-/ra/ pairs, a finding that replicates the stop + /l/ and stop + /r/ findings of Day (1968) and Cutting (1975). In addition, /da/-/ra/ pairs yielded many anomalous fusions: that is, in addition to the cases where listeners wrote down DRA, they also wrote occasional DLA, BLA, and GLA responses. See Day (1968), Cutting (1975), and Cutting and Day (1975) for more specific accounts of this effect.

In general, the results for the six types of fusion fall into three groups when relative onset time is varied. The first group consists of sound localization alone, where the fused percept disintegrates after asynchronies of only 5 msec. The second group consists of psychoacoustic fusion, spectral fusion, and spectral/temporal fusion. When corrections are made for differences in lower asymptote, the three functions are much the same: after asynchronies of 20 and 40 msec, the frequency of fusions tapers off rapidly. Phonetic feature fusion and phonological fusion form a third group, where fusion increases with lead times of 20 to 40 msec, before decreasing with longer lead times.

Consider now the stimulus "units" that appear to be fused in each of these three groups. Sound localization, the only member of Group 1, occurs through cross-correlation of opposite-ear waveforms. In the present study the two signals are speech sounds carried on a glottal source of 100 Hz. Each glottal pulse is thus 10 msec in duration and would serve as a convenient acoustic "unit" to anchor the cross-correlation process. Microstructure differences between contiguous glottal pulses are slight but may serve as an aid in the localization process (Leakey et al., 1958). Onset asynchronies less than 5 msec are apparently surmountable in sound localization, whereas larger differences generally are not. It may be that the binaural hearing mechanisms can integrate glottal pulses conveying identical information if they arrive within 5 msec of one another, in part, because each pulse overlaps by at least half a glottal wavelength. With onset times greater than 5 msec, the opposite-ear pulses that arrive most nearly at the same time are not identical, and microstructure differences may impede localization. From this account, one would predict that a long, continuous steady-state vowel /a/ with identical microstructures within each glottal pulse would always be localizable as a single item regardless of timing differences. Indeed, Broadbent (1955) and Sayers and Cherry (1957) performed demonstrations similar to this.

Group 2 has three members: psychoacoustic fusion, spectral fusion, and spectral/temporal fusion. In psychoacoustic fusion, the second-formant transitions of the opposite-ear stimuli appear to be the "unit" of fusion. Each is 70 msec in duration, and onset differences of 20 to 40 msec are needed before the fused percept disintegrates. Again, the tolerable onset asynchrony appears to be about "half the fused unit," although for this second-level type fusion that unit is several times larger. In psychoacoustic fusion there is competition of opposite-ear information. If this information is not meshed temporally in the right fashion, rivalry will occur and backward masking is the typical result (see Massaro, 1972, 1974). In spectral fusion and in spectral/temporal fusion, on the other hand, there is no competing information. In both, the "units" are the second-formant transition properly aligned with the first-formant transition to yield the percepts /b/, /d/, or /g/. Again, transition durations are the same and, allowing for differences in asymptote, disintegration of the percept appears to occur at about the same point. Here, information about place of articulation from one ear (either as part of the second-formant resonance or as an isolated chirp) is combined with information about manner of production in first-formant transition. In all three fusions, however, the actual fusion appears to be only incidentally linguistic. Transitions of formants are merged, and subsequent analysis of the fused information reveals it to be linguistically labelable as /b/, /d/, or /g/.

Group 3 consists of phonetic feature fusion and phonological fusion. In both, the "units" are linguistic, but in the first the units are phonetic features and in the second they are the phonemes themselves. The increase in fusion over short time intervals, followed by a decrease at longer lead times, separates these fusions from the others. Phonemes and their composite phonetic features can have acoustic manifestations of 20 to 150 msec, depending on the phoneme and on the feature. Thus, it seems to make less sense to talk in terms of "half a unit's duration" as the threshold for the disintegration of the fused percept. Instead, it makes more sense to speak in terms of the time course of processing those features. In an inverted form, the functions in the right-hand panel of Figure 3 look like J-shaped backward-masking functions found in

vision.[5] As noted previously, disruption by masking can be thought of as a reciprocal process to fusion. Processing of linguistic features appears to be disrupted most readily after initial processing of the first-arriving item has taken place; earlier or later arrival of the second item decreases the chance of such interference. The disruption process allows for the possible misassignment of the feature values in the case of phonetic feature fusion; the disruption allows for the possible combination--and in the case of /ra/-leading-/da/ items, the misassignment of temporal order--of the phonemes themselves in phonological fusion.

The three patterns of results, one for each group, suggest at least three types of analysis relevant to speech perception. Each has its own temporal limit within which fusion occurs, and this limit can be thought of as analogous to different types of perceptual "moments" (see Allport, 1968; Efron, 1970). The smallest type of moment lasts up to about 2 to 5 msec, within which time the waveforms of stimuli presented to opposite ears can be meshed (localized). To exceed this limit is to exceed the resolving capacity of the mechanisms involved. An intermediate-sized moment lasts up to perhaps 40 msec and allows the acoustic features of opposing stimuli to merge. Again, to go beyond this range is, generally, to go beyond the system's ability to process (fuse) the discrepancy in stimulus information. A third moment lasts 20 to 80 msec, a time limit that provides maximal opportunity for misassigning certain linguistic features of competing inputs.

It seems likely that these three types of moments reflect processes that occur concurrently and become relevant to the percept according to variations in the stimuli and in the demands placed on the listener in a particular task. It also seems likely that the different sizes of the moments reflect the level at which fusion occurs: the smaller the interval, the lower in the system the fusion occurs and, conversely, the larger the interval, the higher in the system. This general scheme is an extension of that suggested by Turvey (1973), who found that peripheral processes in visual masking occurred over smaller time domains than central processes. Translating the terms peripheral and central to auditory studies of fusion cannot be straightforward, since there is considerably more pathway interaction between the two ears than between the two eyes before events reach a cortical level. Nevertheless, after substituting the more conservative terms lower level and higher level for peripheral and central, the extended analogy is worth pursuing. For example, Turvey (1973) found that in visual masking peripheral processing was characterized by stimulus integration and central processing by stimulus disruption. Auditory fusions of Group 1 (sound localization) and Group 2 (psychoacoustic fusion, spectral fusion, and spectral/temporal fusion) are integrations of opposite-ear stimuli, whereas those of Group 3 (phonetic feature fusion, phonological fusion) appear to occur because of an interruption of speech perception in midprocess and a subsequent misassignment of features.[6]

---

[5] M. T. Turvey, 1974: personal communication.

[6] Turvey (1973) suggests that in vision there may be two types of integrative processes, one integrating energies and the other integrating features. By analogy, Level 1 auditory fusions would appear to be of the first variety, and Level 2 of the second.

An Additional Consideration: Presentation Mode

If lower-level fusions are those characterized by perceptual integration and higher-level fusions by perceptual disruption, presentation mode ought to have a crucial effect on the probability of fused responses and should provide a test of the distinction. Two modes of presentation are of interest: dichotic presentation, the mode used in all experiments in this paper, and a type of binaural presentation. I will define the terms dichotic and binaural slightly differently than do Woodworth (1938:526) and Licklider (1951:1026). They view dichotic listening as a special form of binaural listening. Dichotic presentation, as the term is used here, occurs when Stimulus A is presented to one ear and Stimulus B to the other ear, regardless of whether or not the two items are identical to one another. Binaural presentation, on the other hand, occurs here when Stimuli A and B are combined and both items are presented to both ears. Table 3 summarizes the percent fusion responses for the six fusions as a function of presentation mode. Dichotic scores are means from the present set of experiments, whereas binaural scores stem from logical considerations and from data collected elsewhere.

Fusions of Group 1 (sound localization) and Group 2 (psychoacoustic fusion, spectral fusion, spectral/temporal fusion) reveal a pattern distinctively different from those of Group 3 (phonetic feature fusion and phonological fusion). For the first four fusions, the number of /da/ responses under binaural conditions is slightly greater than or equal to the number of dichotic fusions. For the other two, however, binaural fusions are considerably less frequent than dichotic fusions. The first four could logically occur at a neural level prior to the cortex, or at least prior to linguistic analysis within the cortex. To a great degree, presentation mode is irrelevant here, and acoustic combination may be similar to neural combination (integration)--a straightforward, primarily additive process. The other two fusions, on the other hand, must occur subsequent to linguistic (and cortical) analysis, since mere mixing of the signals inhibits rather than aids in obtaining the desired percept. This occurs presumably because mixing the signals degrades their separate integrities, and the stimuli mask each other effectively before they ever arrive at some central locus for linguistic analysis. Disruption, then, never has a chance to occur because integration has already occurred. In summary, the first four fusions appear to be only incidentally linguistic, whereas the last two are necessarily linguistic.

In addition to providing a framework for the data discussed above, the upper-level/lower-level scheme, as adapted from Turvey (1973), would predict effects of stimulus energy on fusions at these different levels. He found that stimulus energy affected visual masking at a peripheral level but that it had essentially no effect centrally. Here, one would predict that stimulus energy (intensity) would have a relatively large effect on lower-level fusions and a smaller effect on higher-level fusions. Experiment III was designed to test these predictions.

EXPERIMENT III: RELATIVE INTENSITY IN SIX FUSIONS

Method

Three different randomly ordered sequences of dichotic pairs were recorded in a fashion similar to that of Experiment II: one each for psychoacoustic fusion, phonetic feature fusion, and phonological fusion. No sequences were

TABLE 3: Percent "fusion" responses as a function of presentation mode when items have simultaneous onsets and share the same frequency and intensity (F_1 = first formant; F_2 = second formant).

Fusion type and dichotic pair	Dichotic[a]		Binaural[b]
1. Sound localization /da/ + /da/	100 /da/[c]	=	100 /da/[c]
2. Psychoacoustic fusion /ba/ + /ga/	68 /da/	<	81 /da/[d]
3. Spectral fusion /da/ F_1 + /da/ F_2	85 /da/	<	100 /da/[c]
4. Spectral/temporal fusion /da/ without F_2 transition + F_2 transition of /da/	81 /da/	<	100 /da/[c]
5. Phonetic feature fusion /ba/ + /ta/	43 /da/	>	20 /da/[e]
6. Phonological fusion /da/ + /ra/	73 /dra/[f]	>	15 /dra/[f,g]

[a]Dichotic presentation occurs when Stimulus A is presented to one ear and Stimulus B to the other (regardless of whether or not the two items are identical). Data in this column are means found in Experiments II-IV.

[b]Binaural presentation occurs, for the purposes of the present paper, when stimuli are mixed and both are presented to both ears.

[c]These items are physically identical.

[d]Determined by testing with listeners not participating in present study.

[e]Computed from Halwes (1969:75), who used synthetic speech stimuli very similar to those used in the present studies.

[f]The response /dra/ here represents all fusion responses for the stimulus pair /da/ + /ra/, including those in which a different stop consonant or a different liquid was reported. See Day (1968), Cutting (1975), and Cutting and Day (1975) for further details.

[g]Cutting (1975, Experiment II), using synthetic speech stimuli similar to those used here.

prepared for the others since data on these fusions are either irrelevant (sound localization) or readily obtainable elsewhere (spectral fusion and spectral/temporal fusion). Nine relative intensities were used in all three sequences: one stimulus was always the standard 80 dB re 20 $\mu N/m^2$ item, with the other item decreased in intensity by 0, 5, 10, 15, 20, 25, 30, 35, or 40 dB. Sequences for psychoacoustic and phonetic feature fusions consisted of 36 dichotic pairs: (9 relative intensities) × (2 intensity configurations, an 80 dB stimulus on Channel A or on Channel B) × (2 channel assignments). Again, /ba/ and /ga/ ($\underline{b}^1$ and $\underline{g}^1$) were used for psychoacoustic fusion and /ba/ and /ta/ for phonetic feature fusion. The sequence for phonological fusion was exactly twice as long, allowing for the two fusible pairs /ba/-/la/ and /da/-/ra/. Listeners followed the same instructions for each fusion as in Experiment II.

Results and Discussion

Patterns of results for the three types of fusion in question are shown in Figure 4, along with the results for the other fusions adapted from other sources. Again, results will be discussed in terms of the relative intensity levels where significant decreases in fusion first occur, using the same criterion as in Experiment II.

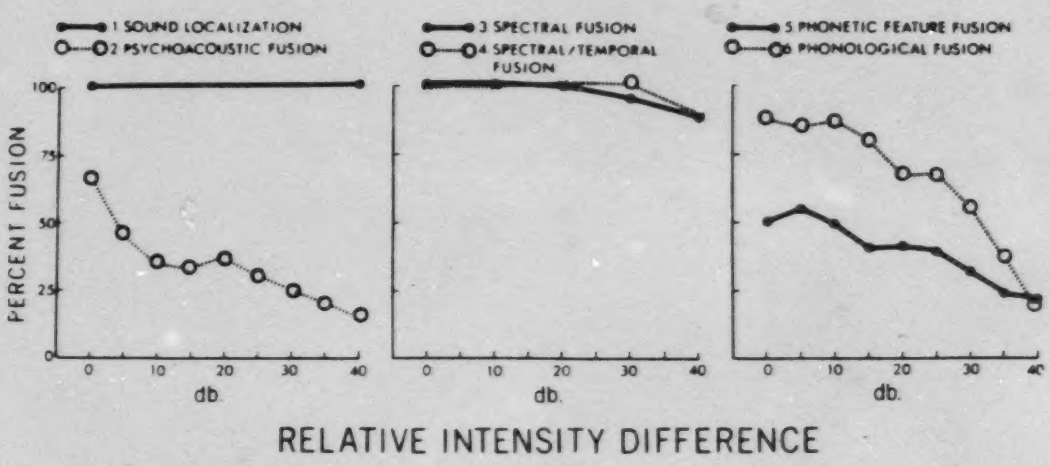


Figure 4: The effect of relative intensity in six fusions. Data from Experiment III for fusions 2, 5, and 6; from Rand (1974) for fusions 3 and 4; and from logical considerations for fusion 1.

For psychoacoustic fusion a significant decrease occurs with a drop of only 10 dB in either stimulus, /ba/ or /ga/ (but see Repp, 1975d). The number of /da/ fusions taper off at a decreased rate thereafter. For both phonetic feature fusion and phonological fusion, on the other hand, significant decreases first occur at 30 dB. In all three types of fusion there was no significant effect of which stimulus in the fusible pair was the most intense. The data plotted for sound localization are hypothetical, but since the fused percept never disintegrates and since binaural interactions can occur over intensity differences greater than 40 dB, a straight line at 100 percent is drawn. The data for spectral fusion and for chirp fusion are adapted from Rand (1974), who used stimuli virtually identical to those in the present study (except his consisted of three formants rather than two).

For a minor replication of the results of Experiment II, compare certain aspects of Figures 3 and 4. Notice that the first data point of each function in

Figure 3 (0-msec lead) represents exactly the same pairs as those in Figure 4 (0 dB). Note further that the probability of fused responses for each fusion at these points is comparable in each study.

Before pursuing the scheme of higher- and lower-level fusions, it is necessary first to reconsider the distinction between those fusions whose stimuli "compete" with one another for the same processor and those whose stimuli do not. Competition, here, is defined as a situation conducive to substantive information loss or alteration. Consider the fusions without competition first. In three of the fusions there is no competition: sound localization (Group 1), spectral fusion, and spectral/temporal fusion (both from Group 2). There is no information loss or alteration in sound localization because the two inputs are identical, and the percept can change only in its perceived locus, not in its identity. There is no information loss in either spectral fusion or spectral/temporal fusion because the acoustic information of opposite-ear stimuli is simply restructured back into the original form from component parts in a straightforward manner. In the other three fusions, however, there is competition. In psychoacoustic fusion (the only other member of Group 2), the second-formant transitions are in the same general space-time region and, as demonstrated in Experiment I, appear to contribute to the fused percept best when they are closest together, apparently enabling them to be perceived as a single formant transition. Information is lost in that /ba/ and /ga/ are no longer heard, and information is altered because both items contribute to the percept /da/. In phonetic feature fusion (Group 3), the stimuli compete for the same limited-channel-capacity linguistic processor in the left hemisphere (see Studdert-Kennedy et al., 1970, 1972). Of two values of the voicing feature and two values of a place-of-articulation feature, only one value of each can typically be processed, while the other is often lost. If they did not belong originally to the same stimulus, a fusion has occurred. Finally, in phonological fusion (Group 3), the stimuli may be processed in opposite hemispheres (Cutting, 1973) with no information loss, but there is considerable information alteration since the two inputs are combined in the most phonologically reasonable fashion to form a single phoneme string. A more detailed comparison of mechanisms thought to underlie the six fusions will be given in the concluding discussion.

Setting aside those fusions in which there is no competition, and hence essentially no effect of intensity with attenuations of as much as 40 dB, one finds that the results of the other three fusions support the higher-level/lower-level distinction discussed earlier. Psychoacoustic fusion, a lower-level process characterized by perceptual integration, is quite sensitive to relatively small attenuations of intensity. On the other hand, phonetic feature fusion and phonological fusion, higher-level processes characterized by perceptual disruption, are relatively insensitive to intensity variation.

The results of Experiments II and III and the additional consideration of presentation mode provide clear evidence for the distinction between auditory processes of Group 2 (psychoacoustic fusion, spectral fusion, and spectral/temporal fusion) and linguistic processes of Group 3 (phonetic feature fusion and phonological fusion). This distinction is similar to that made by Studdert-Kennedy et al. (1972), Pisoni (1973), and Wood (1975), among many others, using very different paradigms. However, evidence thus far for the distinction between Group 1 (sound localization) and Group 2, is less impressive--seen only in the left and center panels of Figure 3. Experiments IV and V are directed at supporting this distinction.

EXPERIMENT IV: RELATIVE FUNDAMENTAL FREQUENCY IN SIX FUSIONS

Method

Six different randomly ordered sequences of dichotic pairs were recorded, one for each type of fusion. Four fundamentals were selected: the standard frequency of 100 Hz, and three others--102, 120, and 180 Hz. Pairs always consisted of one stimulus at 100 Hz, and the other stimulus at any of the four possible fundamentals, yielding relative frequency differences of 0, 2, 20, and 80 Hz. For the sound localization sequence, /da/ was recorded on both channels in a 16-pair sequence: (4 relative fundamentals) × (2 frequency configurations, the 100-Hz item on Channel A or on Channel B) × (2 observations per pair). For the psychoacoustic fusion sequence, /ba/-/ga/ pairs were recorded in a 32-pair sequence: (4 relative frequencies) × (2 frequency configurations) × (2 channel assignments) × (2 observations per pair). Twenty-four item sequences were recorded for both spectral fusion and spectral/temporal fusion. In spectral fusion, the first formant was always held constant at 100 Hz, and in spectral/temporal fusion, the first formant and steady-state segment of the second formant were also always at 100 Hz: frequency variation always took place in the second formant or second-formant transition. The dichotic pairs could yield /ba/, /da/, or /ga/ responses. There were (3 stimulus pairs, those for /ba/, /da/, and /ga/) × (4 relative frequencies) × (2 frequency configurations). Channel assignments were randomized across pairs. For phonetic feature fusion, /ba/-/ta/ pairs were recorded in a 32-pair sequence following the same format as the psychoacoustic fusion sequence, and for phonological fusion a similar 32-pair sequence was recorded for /ba/-/la/ and /da/-/ra/ pairs. Again, listeners followed the same instructions for each fusion as in Experiment II.

Results and Preliminary Discussion

As shown in Figure 5, frequency differences affected only one type of fusion: sound localization. Fusions, the number of one-item responses, plummeted from nearly 100 percent for identically pitched pairs to nearly 0 percent for pairs with only 2-Hz difference between members. Frequency differences had no significant effect on any of the other five types of fusion. Note again the fusion probabilities for pairs with 0-Hz differences are similar to the standard pairs in Experiments II and III.

The results suggest a clear distinction between fusions of Groups 2 and 3 and sound localization of Group 1. A problem arises, however, when one considers that these groups correlate perfectly with the type of response required of the listener. In sound localization, the listener reports whether he heard one item or two; in all other fusions, the listener identifies the item heard. It may be that when frequency varies in these other fusions, the listener could easily report whether one or two items were actually presented, but that since a linguistic response is required, the cues of numerosity are ignored. Experiment V investigates this possibility.

EXPERIMENT V: HOW MANY ITEMS ARE HEARD?

It is clear that sound localization is a very different kind of fusion than the other five, both phenomenologically and in terms of the results of Experiments II and IV. In sound localization the items presented to opposite ears are either integrated into a single percept, or they are not, and the identity of

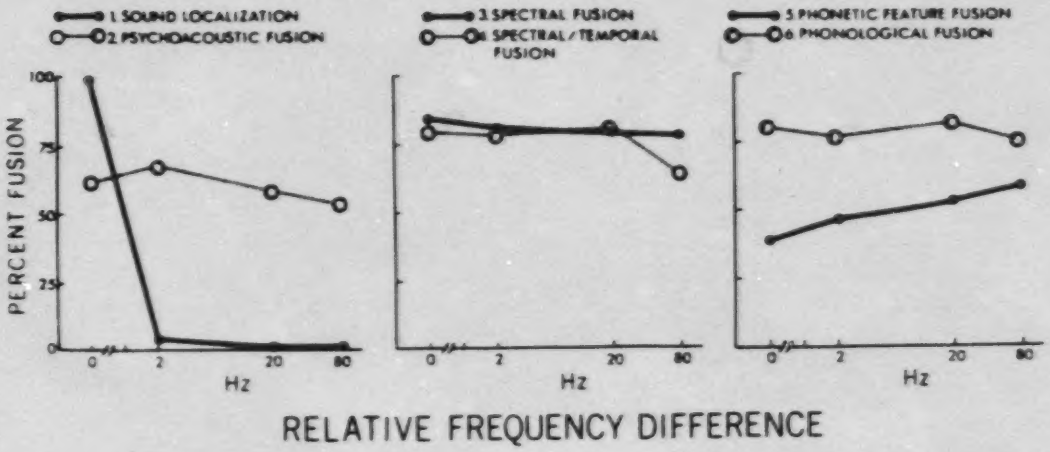


Figure 5: The effect of relative fundamental frequency in six fusions. Data from Experiment IV.

the inputs matters not at all. In all other fusions, by contrast, it may be possible for the listener to be aware that more than one item is presented on a given trial, but to find the fusion response the best label for what he hears. Differences of only 2 Hz convince the listener of the presence of two different items in sound localization. Is this true of other fusions as well? In other words, is it necessary that the items fuse into a single acoustic percept for them to be labeled and judged as a single linguistic percept?

Method

One pair of stimuli was chosen to represent each of the six fusions. They are shown schematically in Figure 1. Pairs were either of the standard form (both items at 100 Hz) or they differed by 2 Hz (one item at 100 Hz and the other at 102 Hz). The standard (80-dB) intensities were used. A sequence of 48 simultaneous-onset pairs was recorded: (6 pairs, one for each fusion) × (2 relative frequencies, 0- or 2-Hz difference) × (2 channel arrangements) × (2 observations per pair). Twenty listeners, ten from the previous experiments and ten others selected according to the same criteria, wrote down 1 or 2, indicating the number of items that they heard on each trial. No practice was given.

Results and Discussion

The results for all six types of fusion are shown in Table 4. For four fusions the number of one-item responses dropped significantly when fundamental frequency varied. These are sound localization (90 percent decrease), psychoacoustic fusion (68 percent), spectral fusion (58 percent), and phonological fusion (41 percent). In the other two the decreases were considerably smaller; only 3 percent for spectral/temporal fusion and 14 percent for phonetic feature fusion.

TABLE 4: Percent one-item responses given to dichotic pairs in Experiment V.

Fusion type and dichotic pair	Stimulus condition	
	Both items at 100 Hz	One item at 100 Hz, other at 102 Hz
1. Sound localization /da/ + /da/	99	9[a]
2. Psychoacoustic fusion /ba/ + /ga/	78	10[a]
3. Spectral fusion /da/ F_1 + /da/ F_2	60	2[a]
4. Spectral/temporal fusion /da/ without F_2 transition + F_2 transition of /da/	15	12
5. Phonetic feature fusion /ba/ + /ta/	22[b]	8
6. Phonological fusion /da/ + /ra/	46	5[a]

[a] Across the two stimulus conditions, differences are significant, $p < .01$, by Wilcoxon matched-pairs signed-ranks test.

[b] Halwes (1969), using similar stimuli, found that listeners reported hearing only one sound when both items shared the same pitch. The difference between his results and those of the present study may be attributable to a contrast effect here induced by mixed presentation of fusible pairs. Halwes blocked pairs of same and different pitches.

Figure 5 and Table 4 demonstrate conclusively that sound localization is different from the other fusions. In those five, the number of items perceived plays no role in the linguistic identity of the fused percepts; in the present experiment, standard pairs may be perceived as a single item most of the time (as in psychoacoustic fusion), a single item about half of the time (as in spectral fusion and phonological fusion), or they may nearly always be perceived as two items (as in spectral/temporal fusion or phonetic feature fusion). The 2-Hz difference (nonstandard) pairs fuse as readily as the standard pairs, yet all nonstandard pairs are perceived as two-item presentations.

In view of the initial formulation of six different fusions in dichotic listening, the most important result here is that spectral fusion and spectral/temporal fusion differ significantly [$T(15) = 0$, $p < .01$] in the number of items

perceived when the to-be-fused stimuli have the same pitch. Respective one-item response frequencies were 60 percent versus 15 percent. A review of Figures 3, 4, and 5 shows no impressive difference between the two fusions: both are moderately insensitive to relative onset-time differences and very insensitive to relative intensity and relative frequency differences. In Table 5, however, one finds support for their separation in the evidence that the fusions are phenomenologically different for the listener.

TABLE 5: Upper limits of interstimulus discrepancies permitting consistent and frequent fusions of /da/ in the six different types of fusion, updating Table 1.

Fusion type	Onset time	Intensity	Frequency
1. Sound localization	<5 msec	--[a]	<2 Hz
2. Psychoacoustic fusion	<40 msec	<10 dB	>80 Hz
3. Spectral fusion	<40 msec	40 dB[b]	>80 Hz
4. Spectral/temporal fusion	<40 msec	40 dB[b]	>80 Hz
5. Phonetic feature fusion	<80 msec	25 dB	>80 Hz
6. Phonological fusion	>80 msec	25 dB	>80 Hz

[a]Intensity differences are not relevant to sound localization as discussed here, since the fused percept never disintegrates with such variation.

[b]Rand (1974), using stimuli very similar to those used in the present study.

GENERAL DISCUSSION: SIX FUSIONS IN SPEECH PERCEPTION

Overview

There are four primary results to be emphasized in the five experiments presented here. First, Experiment I was successful in demonstrating that psychoacoustic fusion is the result of perceptual averaging of optimally similar acoustic features of opposite-ear stimuli. Second, Experiments II-IV replicated the general findings and estimates reported in Table 1, and filled in the empty cells for psychoacoustic fusion and spectral/temporal fusion. These are shown in revised form in Table 5, all based on the syllable /da/. Third, the results of Experiments II-V provided patterns of results that were used to differentiate the six fusions and to arrange them into three groups according to a levels-of-processing analysis. Those levels are discussed below. Fourth, the results of Experiments IV and V suggest that the perception of a single event is not necessary for the assignment of a single linguistic response in the five fusions excluding sound localization.

The results of these experiments and the supporting evidence cited throughout the paper preclude the possibility that the six fusions can be accounted for in terms of a single general mechanism: the large variation in sensitivities to relative onset-time differences and to intensity differences, and the effects of pitch on numerosity versus identity of the fused percepts prevent any suggestions of a simple fusion system. Instead, at least three, perhaps even four, perceptual levels are needed, one each to accomplish the different kinds of perceptual combination.

Level 1: Fusion by Integration of Waveforms

Sound localization is the only fusion to occur at this first and lowest level, and the mechanism involved is one that cross correlates the waveform of opposite-ear stimuli. Three kinds of evidence separate this fusion from the other five. First, as shown in Experiment II, no other fusion is as sensitive to differences in relative onset time. Onset differences of 4 and 5 msec are sufficient to inhibit fusion, nearly an order of magnitude less than those intervals necessary for other fusions to disintegrate. Second, extreme sensitivity to relative frequency differences in two speech sounds is very clear from Experiment IV; whereas frequency differences do not affect other fusions. Third, on logical grounds alone, sound localization is unique because it is the only fusion based on number of percepts rather than on their identity. See Deatherage (1966) for an account of the physiology of the binaural system, and Sayers and Cherry (1959) for an indication of interactions involved in sound localization.

Level 2: Fusion by Integration of Acoustic Features

Three fusions appear to occur at this second level: psychoacoustic fusion, spectral fusion, and spectral/temporal fusion. Evidence for their common allocation comes from five sources. First, each is moderately sensitive to differences in relative onset time, yielding markedly similar patterns especially when corrections are made for differential floor effects. Experiment II found that each withstood temporal differences of between 20 and 40 msec without marked disintegration of the fused percept. Second, also stemming from Experiment II, the shape of the functions as onset interval increases is quite different from those of phonetic feature fusion and phonological fusion, suggesting an entirely different process. Third, none of these three fusions is dependent on prior perception of the dichotic inputs as a single event. Experiments IV and V taken together demonstrate that frequency differences have no effect on the probability of fusion responses, but that for psychoacoustic fusion and spectral fusion frequency differences do significantly affect the numbers of items perceived to occur on any given trial. Fourth, although this third stipulation is also true for phonetic feature fusion and phonological fusion, these three fusions differ from the two remaining fusions with respect to the importance of presentation mode. As shown in Table 3, there are as many, or more, "fusions" for these three phenomena when the items are presented binaurally as when they are presented dichotically. (Phonetic feature fusion and phonological fusion, on the other hand, show the reverse trend.) These data and those for similar phenomena reported elsewhere (Pisoni and McNabb, 1974; Repp, 1975a) suggest that this second level must be prephonetic. Fifth, relative stimulus intensity plays an important role in the only one of these fusions that occurs through dichotic competition--psychoacoustic fusion. Sensitivity to relative energy levels is indicative of lower-level processing in audition.

Although they occur at the same perceptual level, these three fusions are separate phenomena. Psychoacoustic fusion, a general phenomenon most dramatically represented by the fusion of /ba/ and /ga/ into /da/, occurs through the perceptual averaging of similar but slightly discrepant information presented to opposite ears. The same averaging could be accomplished using synthetic steady-state vowels but with slightly different vowel color: a vowel of midcolor between the two inputs is easily perceived. No such averaging occurs in spectral

or spectral/temporal fusion since there is no discrepant information competing between the two ears.[7]

Spectral fusion and spectral/temporal fusion differ strikingly in perceptual "appearance" to the listener. When fundamental frequencies of the to-be-fused stimuli are the same, listeners generally report hearing only one item in spectral fusion but two items in spectral/temporal fusion. This remarkable fact is sufficient to consider them separate phenomena, and requires further consideration of spectral/temporal fusion.

In spectral/temporal fusion, the second-formant transition is "heard" in two forms: one as part of a speech syllable giving it its identity as /ba/, /da/, or /ga/, and the other as a brief glissando or chirp. How does this dual perception come about? In part, one must appeal to dual perceptual systems of speech and nonspeech (Day and Cutting, 1971; Mattingly et al., 1971; Day and Bartlett, 1972; Day, Bartlett, and Cutting, 1973). The brief chirp appears to be fused (integrated) with the transitionless stimulus at Level 2 and then identified by a "speech processor." The information in this chirp also appears to remain in a relatively raw acoustic form--a brief acoustic blip against the background of a continuous periodic speech sound. In spectral fusion, by contrast, there is no brief signal to establish this figure-ground relationship. Logically, however, there are three possible percepts in the spectral/temporal fusion of /da/, as shown in Figure 1: the speech sound /da/ and the chirp, the two sounds that are actually heard, and the transitionless /da/, which is not heard. Why not? The following account is somewhat complex, but appears to explain the phenomenon and an additional anomaly.

The transitionless /da/ stimulus is about 85 percent identifiable as /ba/. As noted earlier, it is identified as /ba/ presumably because the harmonics of the first-formant transition mimic the absent second-formant transition and mimic it in such a fashion as to be appropriate for /b/. Thus, at some level, this /ba/ competes with a redintegrated /da/. The redintegrated /da/, like the original stimulus, is a highly identifiable, hypernormal (Mattingly, 1972) item. It competes with the considerably weaker /ba/, readily identifiable but without a prominent transitional cue typical of synthetic speech syllables. The /da/ easily "wins" in this mismatch, and is perceived instead of /ba/. Data supporting this account show that fewer /ga/ spectral/temporal fusions occur than /da/ fusions given the appropriate stimuli (Nye, Nearey, and Rand, 1974; see also Experiment II, present paper). Here the transitionless /ga/ (again perceived as /ba/) competes with the redintegrated /ga/. The result is that fewer /ga/ percepts arise and a number of /da/ responses are reported instead. This /da/-for-/ga/ substitution may arise from the psychoacoustic fusion of the transitionless /ga/ (perceived as /ba/) and the redintegrated /ga/.

If psychoacoustic fusion, spectral fusion, and spectral/temporal fusion occur at the same level of processing in the auditory system, it should be possible, as suggested above, for them to interact directly to create composite

---

[7] The results of Perrott and Barry (1969), showing fusion of sine waves whose frequencies differ beyond the range of binaural beats, might be thought to be a form of psychoacoustic fusion, but since their task required detection of one versus two signals, and not the "identification" of the signals, their phenomenon appears to be one more closely allied to sound localization.

fusions. Indeed, this appears to be possible. Consider two composite fusions: the first a combination of psychoacoustic fusion and spectral fusion and the second a combination of psychoacoustic fusion and spectral/temporal fusion. In the first case, if the syllable /ba/ is presented to one ear and the second formant of /ga/ to the other, the listener can easily hear /da/, and probably with approximately the same probability as she heard it when the stimuli were /ba/ and /ga/. Berlin, Porter, Lowe-Bell, Berlin, Thompson, and Hughes (1973) have performed experiments similar to this, and from their results the fusion in this situation seems likely. In the second case, the syllable /ba/ might be presented to one ear and the /ga/ second-formant transition to the other. In this situation, the listener may report hearing /da/ plus chirp. Pilot research supports the likelihood of these two composite fusions.

Level 3: Fusion by Disruption and Recombination of Linguistic Features

Phonetic feature fusion and phonological fusion can be separated from the other fusions by three findings. First, the functions revealed when relative onset time is varied are unique: for both phenomena, fusions increase slightly with small onset-time asynchronies only to decrease after onsets of greater than about 40 to 80 msec. This nonlinearity suggests a multistage process similar, perhaps, to that proposed by Turvey (1973). Fusion occurs best, it would seem, only after the first-arriving stimulus has been partially processed, features extracted from it, but immediately disrupted by the arrival of the second item. Thus, these fusions are perceptual confusions resulting from the misassignment of linguistic features. Second, these fusions are only moderately sensitive to intensity differences between the two stimuli, and provide sharp contrast to psychoacoustic fusion, the only other of these six phenomena that occurs through perceptual competition of similar information presented to both ears. Insensitivity to relative energies is characteristic of higher-level processes (see Turvey, 1973, for a visual parallel). Third, as cited previously in this discussion, phonetic feature fusion and phonological fusion are the only fusions to suffer when presentation mode is charged from dichotic to binaural. This occurs presumably because the mixing of the signals degrades their intelligibility, and the stimuli mask one another through integration before disruption can ever take place.

These are the two fusions in which the actual combinative process is necessarily linguistic. With the possible exception of spectral/temporal fusion, the responses of the listeners for other fusion phenomena are only incidentally linguistic. Although the results presented here do not distinguish phonetic feature fusion from phonological fusion, several other results and logical considerations may warrant separate linguistic levels for the two.

Levels 3 and 4? Empirical findings and several logical considerations distinguish the two language-based fusions. In Figure 3 it appears that phonological fusion is only slightly more tolerant of lead time differences than in phonetic feature fusion. This apparent similarity may be misleading. Day (1970b), Day and Cutting (1970), and Cutting (1975) have shown that when longer and more complex speech stimuli are used, such as one- and two-syllable words, tolerance to onset-time asynchronies can increase from about 80 msec to at least 150 msec, considerably beyond any consistent effects of phonetic feature fusion. Since higher-level fusions typically allow for greater tolerance to relative onset differences, phonetic feature fusion might be thought to occur at Level 3 and phonological fusion at a new level, Level 4. A second finding that might

support the separation of the two linguistic fusions is that several previous studies in phonological fusion (Cutting, 1975; Cutting and Day, 1975) have found that pairs analogous to /da/-leading-/ra/ fuse more readily than pairs like /ra/-leading-/da/. Such asymmetry does not occur for phonetic feature fusion.

In addition, logical considerations support the separation of the two phenomena. Phonological constraints, those dictating the logic of contiguity for phonemes in a given language, are higher-level language constraints than are phonetic feature analyses (see, among others, Kirstein, 1973; Studdert-Kennedy, 1974, in press). For example, whereas phonetic features are almost universal across all languages, phonologies are language specific. In English, liquids (/l/ and /r/) cannot precede stop consonants (/b/, /g/, /p/, and /k/, for example) in initial position, but stops readily come before liquids. This appears to account for the fact that given a dichotic pair such as BANKET/LANKET the listener rarely reports hearing LBANKET. Instead, he often hears BLANKET.

Another consideration is also important and may separate phonological fusion from all other fusion phenomena. Day (1970a, 1974), Cutting and Day (1975), and to a lesser extent Cutting (1975), found that there are marked individual differences in the frequency of phonological fusions across different subjects. Some individuals fuse very frequently, others fuse relatively less often, and few individuals fuse at rates in between these two modes. Moreover, these differences correlate with those found on other tasks with the same stimuli and on tasks involving very different stimuli (Day, 1970a, 1973, 1974). Preliminary results suggest that such radical and systematic differences may not occur elsewhere in the six fusions.

At least one additional consideration, however, supports the notion that the two linguistic fusions do indeed occur at the same level. Hofmann (1967) and Menyuk (1972) have suggested that clusters of phonemes, including stop-liquid clusters, may be more parsimoniously described as single underlying phonemes with their own unique phonetic features (see also Devine, 1971). The results of Cutting and Day (1975, Experiment IV) appear to support this conclusion in that certain aspects of the dichotic presentation may mimic certain phonetic feature values and contribute substantially to the perception of a stop /l/ cluster. Thus, "blending" of phonetic features might account for both phonetic feature fusion and phonological fusion. In summary, then, data and logical considerations may suggest a separation of the two fusions, but the separation cannot yet be affirmed.

SUMMARY AND CONCLUSION

Fusion is not a single phenomenon in speech perception, but many. Six dichotic fusions were considered, and five of them are distinctive in that the fused percept differs from either of the two inputs. The robustness of these phenomena was measured against variation in three parameters: relative onset time of the two stimuli, relative intensity of the stimuli, and their relative frequency. Results, gathered here using the same subjects and essentially the same stimulus repertory for each fusion, agree with previously published accounts or, where there are no prior data, fit nicely into the scheme of upper- and lower-level fusions developed in this paper. The various fusions cannot occur at a single perceptual level: at least three, perhaps four, levels are needed.

Fusion, on the one hand, and rivalry and masking, on the other, allow reciprocal glances at the same phenomena. The levels of perceptual processing developed here for audition are quite similar to those developed elsewhere in audition (Studdert-Kennedy, 1974; Wood, 1974, 1975) and also those developed in vision (Turvey, 1973). With the exception of Julesz (1971), most research in both modalities concerning itself with stages of processing has used rivalry/masking paradigms. The findings in the present paper suggest that fusion paradigms can also be used to probe the speech-processing system.

REFERENCES

Ades, A. E. (1974) Bilateral component in speech perception? J. Acoust. Soc. Am. 56, 610-616.

Allport, D. A. (1968) Phenomenal simultaneity and the perceptual moment hypothesis. Brit. J. Exp. Psychol. 59, 395-406.

Angell, J. R. and W. Fite. (1901) The monaural localization of sound. Psychol. Rev. 8, 225-246.

Berlin, C. E., R. J. Porter, S. S. Lowe-Bell, H. L. Berlin, C. L. Thompson, and L. F. Hughes. (1973) Dichotic signs of the recognition of speech elements in normals, temporal lobectomees, and hemispherectomees. IEEE Trans. Audio Electroacoust. AU-21, 189-195.

Bever, T. G. and R. J. Chiarello. (1974) Cerebral dominance in musicians and nonmusicians. Science 195, 537-539.

Bilsen, F. A. and J. L. Goldstein. (1974) Pitch of dichotically delayed noise and its possible spectral basis. J. Acoust. Soc. Am. 55, 292-296.

Blechner, M. J., R. S. Day, and J. E. Cutting. (in press) Processing two dimensions of nonspeech stimuli: The auditory phonetic distinction reconsidered. J. Exp. Psychol.: Human Perception and Performance. [Also in Haskins Laboratories Status Report on Speech Research SR-42/43 (1975), 221-232.]

Broadbent, D. E. (1955) A note on binaural fusion. Quart. J. Exp. Psychol. 7, 46-47.

Broadbent, D. E. and P. Ladefoged. (1957) On the fusion of sounds reaching different sense organs. J. Acoust. Soc. Am. 29, 708-710.

Cherry, E. C. and W. K. Taylor. (1954) Some further experiments upon the recognition of speech with one ear and with two ears. J. Acoust. Soc. Am. 26, 554-559.

Cooper, F. S. and I. G. Mattingly. (1969) Computer-controlled PCM system for investigation of dichotic speech perception. J. Acoust. Soc. Am. 46, 115(A).

Cramer, E. M. and W. H. Huggins. (1958) Creation of pitch through binaural interaction. J. Acoust. Soc. Am. 30, 413-417.

Cullen, J. K., C. L. Thompson, L. F. Hughes, C. E. Berlin, and D. S. Samson. (1974) The effects of varied acoustic parameters on performance in dichotic speech perception tasks. Brain Lang. 1, 307-322.

Cutting, J. E. (1972) A preliminary report on six fusions in auditory research. Haskins Laboratories Status Report on Speech Research SR-31/32, 93-107.

Cutting, J. E. (1973) Levels of processing in phonological fusion (doctoral dissertation, Yale University). Dissertation Abstracts International 34, 2332B (University Microfilms No. 73-25191). [Also published in Haskins Laboratories Status Report on Speech Research SR-34 (1973), 1-53.]

Cutting, J. E. (1974) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys. 16, 601-612.

Cutting, J. E. (1975) Aspects of phonological fusion. J. Exp. Psychol.: Human Perception and Performance 1, 105-120.

Cutting, J. E. (in press) The magical number two and the natural categories of speech and music. In Tutorial Essays in Psychology, ed. by N. S. Sutherland. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.). [Also in Haskins Laboratories Status Report on Speech Research SR-42/43 (1975), 189-219.]

Cutting, J. E. and R. S. Day. (1975) The perception of stop-liquid clusters in phonological fusion. J. Phonetics 3, 99-113.

Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.

Cutting, J. E., B. S. Rosner, and C. F. Foard. (in press) Perceptual categories for musiclike sounds: Implications for theories of speech perception. Quart. J. Exp. Psychol.

Darwin, C. J. (1971) Dichotic backward masking of complex sounds. Quart. J. Exp. Psychol. 23, 386-392.

Day, R. S. (1968) Fusion in dichotic listening (doctoral dissertation, Stanford University). Dissertation Abstracts International (1969) 29, 2649B (University Microfilms No. 69-211).

Day, R. S. (1970a) Temporal-order judgments in speech: Are individuals language-bound or stimulus-bound? Haskins Laboratories Status Report on Speech Research SR-21/22, 71-87.

Day, R. S. (1970b) Temporal-order perception of a reversible phoneme cluster. J. Acoust. Soc. Am. 48, 95(A).

Day, R. S., (1973) Individual differences in cognition. Paper presented at the 13th meeting of the Psychonomic Society, St. Louis, Mo., November.

Day, R. S. (1974) Differences in language-bound and stimulus-bound subjects in solving word-search puzzles. J. Acoust. Soc. Am. 55, 412(A).

Day, R. S. and J. C. Bartlett. (1972) Separate speech and nonspeech processing in dichotic listening? J. Acoust. Soc. Am. 51, 79(A).

Day, R. S., J. C. Bartlett, and J. E. Cutting. (1973) Memory for dichotic pairs: Disruption of ear performance by the speech/nonspeech distinction. J. Acoust. Soc. Am. 53, 358(A).

Day, R. S. and J. E. Cutting. (1970) Levels of processing in speech perception. Paper presented at the 10th meeting of the Psychonomic Society, San Antonio, Tex., November.

Day, R. S. and J. E. Cutting. (1971) What constitutes perceptual competition in dichotic listening? Paper presented at the annual meeting of the Eastern Psychological Association, New York, April.

Deatherage, B. (1966) An examination of binaural interaction. J. Acoust. Soc. Am. 39, 232-249.

Delattre, P. C., A. M. Liberman, and F. S. Cooper. (1955) Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769-773.

Deutsch, D. (1975a) Musical illusions. Sci. Am. 233(4), 92-105.

Deutsch, D. (1975b) Two-channel listening to musical scales. J. Acoust. Soc. Am. 57, 1156-1160.

Deutsch, D. and P. L. Roll. (in press) Separate "what" and "where" decision mechanisms in processing a dichotic tonal sequence. J. Exp. Psychol.: Human Perception and Performance.

Devine, A. M. (1971) Phoneme or cluster: A critical review. Phonetica 24, 65-85.

Efron, R. (1970) The relationship between the duration of a stimulus and the duration of a perception. Neuropsychologia 8, 37-55.

Fourcin, A. J. (1962) An aspect of the perception of pitch. In Proceedings of the Fourth International Congress of Phonetic Sciences, ed. by A. Sovijarvi and P. Aalto. (The Hague: Mouton), pp. 355-359.

Fry, D. P. (1956) Perception and recognition in speech. In For Roman Jakobson, ed. by M. Halle, H. G. Lunt, and C. H. Schoonveld. (The Hague: Mouton), pp. 169-173.

Groen, J. J. (1964) Super- and subliminal binaural beats. Acts Otolaryngol. 57, 224-230.

Guttman, N. and B. Julesz. (1963) Lower limits of auditory periodicity analysis. J. Acoust. Soc. Am. 35, 610.

Haggard, M. (1975) Asymmetrical analysis of stimuli with dichotically split formant information. (Speech Perception, Report on Speech Research in Progress, Psychology Department, The Queen's University of Belfast) Series 2, no. 4, 11-19.

Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech (doctoral dissertation, University of Minnesota). Dissertation Abstracts International (1970) 31, 1565B (University Microfilms No. 70-15736). [Issued as Supplement to the Haskins Laboratories Status Report on Speech Research.]

Hofmann, T. R. (1969) Initial clusters in English. Quarterly Progress Report (Research Laboratories of Electronics, MIT) 84, 263-274.

Houtsma, A. T. M. and J. L. Goldstein. (1972) The central origin of the pitch of complex tones: Evidence from musical interval recognition. J. Acoust. Soc. Am. 51, 520-529.

Huggins, A. W. F. (1964) Distortion of the temporal pattern of speech: Interruption and alternation. J. Acoust. Soc. Am. 36, 1055-1064.

Huggins, A. W. F. (1974) On the perceptual integration of dichotically alternating pulse trains. J. Acoust. Soc. Am. 56, 939-943.

Huggins, A. W. F. (1975) Temporally segmented speech. Percept. Psychophys. 18, 149-157.

Jeffress, L. A. (1972) Binaural signal detection: Vector theory. In Foundations of Modern Auditory Theory, Vol. 2, ed. by J. V. Tobias. (New York: Academic Press), pp.349-368.

Julesz, B. (1971) Foundations of Cyclopean Perception. (Chicago: University of Chicago Press).

Kaufman, L. (1963) On the spread of suppression and binocular rivalry. Vision Res. 3, 401-415.

Kaufman, L. (1974) Sight and Mind. (New York: Oxford University Press).

Kirstein, E. F. (1973) The lag effect in dichotic speech perception. Haskins Laboratories Status Report on Speech Research SR-35/36, 81-106.

Kubovy, M., J. E. Cutting, and R. M. McGuire. (1974) Hearing with the third ear: Dichotic perception of a melody without monaural familiarity cues. Science 186, 272-274.

Leakey, D. M., B. M. Sayers, and E. C. Cherry. (1958) Binaural fusion of low- and high-frequency sounds. J. Acoust. Soc. Am. 30, 222.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 631-661.

Licklider, J. C. R. (1951) Basic correlates of the auditory stimulus. In Handbook of Experimental Psychology, ed. by S. S. Stevens. (New York: Wiley), pp. 985-1039.

Licklider, J. C. R., J. C. Webster, and J. M. Hedlun. (1950) On the frequency limits of binaural beats. J. Acoust. Soc. Am. 22, 468-473.

Linden, A. (1964) Distorted speech and binaural speech resynthesis tests. Acta Otolaryngol. 58, 32-48.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.
Locke, S. and L. Kellar. (1973) Categorical perception in a nonlinguistic mode. Cortex 9, 355-369.
Marslen-Wilson, W. D. (1975) Sentence perception as an interactive parallel process. Science 189, 226-228.
Massaro, D. W. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.
Massaro, D. W. (1974) Perceptual units in speech recognition. J. Exp. Psychol. 102, 199-208.
Mattingly, I. G. (1972) Speech cues and sign stimuli. Am. Scient. 60, 327-337.
Mattingly, I. G., A. M. Liberman, A. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.
Matzker, J. (1959) Two new methods for the assessment of central auditory function in cases of brain disease. Ann. Otol. Rhinol. Laryngol. 68, 1185-1197.
Menyuk, P. (1972) Clusters as single underlying consonants: Evidence from children's productions. In Proceedings of the Seventh International Congress of Phonetic Sciences, ed. by A. Rigault and R. Charbonneau. (The Hague: Mouton), pp. 1161-1165.
Mills, A. W. (1972) Auditory localization. In Foundations of Modern Auditory Theory, Vol. 2, ed. by J. V. Tobias. (New York: Academic Press), pp. 303-348.
Nearey, T. M. and A. G. Levitt. (1974) Evidence for spectral fusion in dichotic release from upward spread of masking. Haskins Laboratories Status Report on Speech Research SR-39/40, 81-90.
Noorden, L. P. A. S. van (1975) Temporal Coherence in the Perception of Tone Sequences. (Eindhoven, Holland: Instituut voor Perceptie Onderzoek).
Nye, P. W., T. M. Nearey, and T. C. Rand. (1974) Dichotic release from masking: Further results from studies with synthetic speech stimuli. Haskins Laboratories Status Report on Speech Research SR-37/38, 123-137.
Oster, G. (1973) Auditory beats and the brain. Sci. Am. 229(4), 94-103.
Perrott, D. R. and S. H. Barry. (1969) Binaural fusion. J. Audit. Res. 9, 263-269.
Perrott, D. R. and L. F. Elfner. (1968) Monaural localization. J. Audit. Res. 8, 185-193.
Perrott, D. R. and M. A. Nelson. (1969) Limits for the detection of binaural beats. J. Acoust. Soc. Am. 46, 1477-1481.
Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.
Pisoni, D. B. and S. D. McNabb. (1974) Dichotic interaction of speech sounds and phonetic feature processing. Brain Lang. 1, 351-362.
Rand, T. C. (1974) Dichotic release from masking for speech. J. Acoust. Soc. Am. 55, 678-680.
Repp, B. H. (1975a) Dichotic forward and backward "masking" between CV syllables. J. Acoust. Soc. Am. 57, 483-496.
Repp, B. H. (1975b) Dichotic masking of consonants by vowels. J. Acoust. Soc. Am. 57, 724-735.
Repp, B. H. (1975c) Distinctive features, dichotic competition, and the encoding of stop consonants. Percept. Psychophys. 17, 231-242.
Repp, B. H. (1975d) Perception of dichotic place contrasts. Manuscript to be submitted for publication.

Sayers, B. M. and E. C. Cherry. (1957) Mechanism of binaural fusion in the hearing of speech. J. Acoust. Soc. Am. 29, 973-987.

Shankweiler, D. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Smith, B. A. and D. M. Resnick. (1972) An auditory test for assessing brain stem integrity: Preliminary report. Laryngoscope 82, 414-424.

Studdert-Kennedy, M. (1974) The perception of speech. In Current Trends in Linguistics, ed. by T. A. Sebeok. (The Hague: Mouton). [Also in Haskins Laboratories Status Report on Speech Research SR-23 (1970), 15-48.]

Studdert-Kennedy, M. (in press) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic Press). [Also in Haskins Laboratories Status Report on Speech Research SR-39/40 (1974), 1-52.]

Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech. J. Acoust. Soc. Am. 48, 579-594.

Studdert-Kennedy, M., D. P. Shankweiler, and D. B. Pisoni. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. Cog. Psychol. 3, 455-466.

Studdert-Kennedy, M., D. P. Shankweiler, and S. Schulman. (1970) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. Acoust. Soc. Am. 48, 599-602.

Tallal, P. and M. Piercy. (1974) Developmental aphasia: Rate of auditory processing and selective impairment of consonant perception. Neuropsychologia 12, 83-93.

Tallal, P. and M. Piercy. (1975) Developmental aphasia: The perception of brief vowels and extended stop consonants. Neuropsychologica 13, 69-73.

Thurlow, W. R. and L. F. Elfner. (1959) Pure-tone cross-ear localization effects. J. Acoust. Soc. Am. 31, 1606-1608.

Tobias, J. V. (1972) Curious binaural phenomena. In Foundations of Modern Auditory Theory, Vol. 2, ed. by J. V. Tobias. (New York: Academic Press), pp. 464-486.

Turvey, M. T. (1973) On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. Psychol. Rev. 80, 1-52.

Wood, C. C. (1974) Parallel processing of auditory and phonetic information in speech discrimination. Percept. Psychophys. 15, 501-508.

Wood. C. C. (1975) Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. J. Exp. Psychol.: Human Perception and Performance 1, 3-20.

Wood, C. C., W. R. Goff, and R. S. Day. (1971) Auditory evoked potentials during speech perception. Science 173, 1248-1251.

Woodworth, R. S. (1938) Experimental Psychology. (New York: Holt).

Initial Phonemes Are Detected Faster in Spoken Words than in Spoken Nonwords

Philip Rubin,* M. T. Turvey,* and Peter van Gelder+

ABSTRACT

In two experiments, subjects monitored sequences of spoken consonant-vowel-consonant words and nonwords for a specified initial phoneme. In Experiment I the target-carrying monosyllables were embedded in sequences in which the monosyllables were all words or all nonwords. The possible contextual bias of Experiment I was minimized in Experiment II through a random mixing of target-carrying nonwords only in the final consonant, for example, /bit/ vs. /bip/. In both experiments, subjects detected the specified consonant /b/ significantly faster when it began a word than when it began a nonword. One interpretation of this result is that in speech perception lexical information is accessed before phonological information. This interpretation was questioned and preference was given to the view that the result reflected processes subsequent to perception: words become available to awareness faster than nonwords and therefore provide a basis for differential responding that much sooner.

It is commonplace to conceptualize the process of pattern identification as a hierarchically organized sequence of operations that maps the structured energy at the receptors onto increasingly more abstract representations. In its most simplistic form, this conception characterizes the "conversation" between representations as unidirectional, that is, a more abstract representation is constructed with reference to a less abstract representation, but not vice versa. There are, however, a number of curious results that question the integrity of this characterization. By way of example, a briefly exposed and masked letter is recognized more accurately when part of a word than when part of a nonword (Reicher, 1968; Wheeler, 1970). Other, related results suggest that this is a fairly general phenomenon. Thus, detection of an oriented line is significantly better when it is part of a briefly exposed, and masked, unitary picture of a well-formed three-dimensional object than when it is a part of a picture portraying a less well-formed, and flat, arrangement of lines (Weisstein and Harris, 1974). As revealed in the work of Biederman and his colleagues (Biederman, 1972; Biederman, Glass, and Stacy, 1973), this facilitation of "feature" detection by object context is matched by a

---

*Also University of Connecticut, Storrs.

+University of Connecticut Health Center, Farmington.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

facilitation of "object" detection by scene context: an object is more accurately and rapidly identified when part of a briefly exposed real-world scene than when it is part of a jumbled version of that scene, exposed equally briefly.

The present paper reports two experiments that were conducted to determine whether there are phenomena of speech perception analogous to those just described. In several recent experiments, a latency-of-detection task has been used to explore characteristics of speech processing. A case in point is the research of Foss and Swinney (1973) that demonstrated that two-syllable word targets were detected faster than their one-syllable counterparts, and that these in turn were detected faster than individual phonemes. Observations of this kind have motivated legitimate reservations about the relevance of the detection task to the analysis of perceptual stages. We will echo these reservations in our discussion. For the present, however, we draw attention to an important difference between the visual experiments described above and the speech experiments typified by Foss and Swinney (1973) (cf. Savin and Bever, 1970; McNeill and Lindig, 1973). The speech experiments have looked at differences in detection latencies for different kinds of targets (for example, phoneme, syllable). By way of contrast, the visual experiments have held the target type constant and varied the structure in which it is embedded. The question of interest has been the effect different structures have on the detection of their constituent elements. It is this question that provides the point of departure for our experiments.

There are several intimations that the global structure of a speech event significantly influences one's identification or detection of lower-order aspects of the speech signal. For example, it has been reported (Hadding-Koch and Studdert-Kennedy, 1964; Studdert-Kennedy and Hadding, 1973) that judgments about the final movements of a pitch contour--precisely, whether it rises or falls--are that it rises if the total contour is perceived as a question (even if the contour in fact has a final fall), and that it falls if the total contour is perceived as a statement (even if the contour in fact has a final rise). The present experiments examine phoneme targeting in words and nonwords, focusing on the initial phoneme. A demonstration that the detection of an initial phoneme of an utterance is affected by the lexical/semantic value of the utterance would dramatize the influence of holistic and higher-order properties on the detection of speech components.

EXPERIMENT I

Method

Subjects. The subjects were six male and nine female undergraduates at the University of Connecticut. The subjects participated to receive experimental credit in a course on introductory psychology.

Materials and apparatus. Each subject received three blocks of trials, each block consisting of 16 sequences of consonant syllables. Sequences contained either six monosyllabic words or six monosyllabic nonwords. Within a block there appeared an equal number of word and nonword sequences. Each sequence contained exactly one target syllable beginning with the phoneme /b/ or the phoneme /s/, occurring at a point between the second and the fifth

syllable, with equal probabilty for each position in the sequence. Within a block there was an equal number of words and nonwords beginning with /b/ and /s/, and across the block there apeared an equal number of words and nonwords for each position in a sequence. The distinction between lexicon membership and nonmembership--that is, the word/nonword difference--was based on a change in the final consonant. No syllable, target or nontarget, contained a /b/ or /s/ in any position other than the initial position. Phonemes that are highly confusable with the target phonemes (e.g., /f/, /v/, /p/) did not appear in any syllable, target or nontarget. Target syllables were constructed so that the target was followed by different vowel phonemes; and to control for pronounceability, each of the different vowels followed the /b/ and /s/ targets an equal number of times. Table 1 presents some sample sequences. Items in this table are not given their phonetic spellings, but are presented in a form that makes clear the difference between words and nonwords. Presentation by block was either to the left ear, right ear, or both ears. A subject was presented with one of three ordered block sequences: (1) left, right, binaural; (2) right, binaural, left; (3) binaural, left, right. Presentation of particular targets in a sequence was also counterbalanced.

TABLE 1: Experiment I: Sample test sequences.[a]

JUT	LEG	SIN	RUG	WELL	RUN
MEG	GEEL	NUCK	HAEN	BAL	HIG
KEEJ	NUG	LAN	NAEN	SIM	DAJ
COME	BAT	LAG	TELL	TIN	GUM

[a]Target items are underlined.

The speech was recorded at a normal speaking rate by a male speaker on one channel of an Ampex tape recorder. The speech waveform was then digitized and edited using the Haskins Laboratories pulse-code-modulation (PCM) system (Cooper and Mattingly, 1969). In the case of syllables beginning with the phoneme /s/, onset was standardized by starting sampling 100 msec before the start of the vowel. The recording of test tapes involved converting the digitized waveform samples to analog form. Using a Crown 800 tape recorder, test stimuli were recorded on one track of a test tape. The average duration of stimuli, both words and nonwords, was 450 msec. On the second track, a 500-msec, 500-Hz tone appeared coincident with the onset of target items. Syllables were separated by a 1-sec interstimulus interval and sequences of syllables were separated by 5 sec. The 500-Hz tone was used to start a timer in a Data General Nova computer, which involved the computer sampling the signal from track two of the presentation tape through an analog-to-digital converter. When a target item (and its coincident tone) was presented, the real-time clock in the computer was started and the button-push of a subject stopped the clock, thus giving the reaction time of the subject. Reaction times were printed out on a ASR-33 teletype after each block. Reaction times greater than 1500 msec were considered to be errors. The ear of presentation was controlled by feeding the channel-one output of a Sony TC-100 tape recorder

through a mixer that put the signal in either the left, right, or both speakers of two sets of Superex headphones--one set for the subject, and one set for the experimenter to monitor the experiment.

Procedure. Subjects were told that they were going to hear sequences of monosyllables--both nonsense words and real words. Examples were then given both of syllables and sequences of syllables. Subjects were instructed to press a key as fast as possible whenever they heard a word or nonsense word that began with the sound /b/ or /s/. Further examples were given. Subjects were also instructed that their performance was being monitored, and that they should continue targeting even if they made an error. They then heard a practice block of eight sequences to familiarize them with the task. Five seconds before each block, the word "ready" (recorded on track one of the test tape) was presented, to prepare the subjects for the beginning of the block. Before the start of each block, subjects were informed of the ear of presentation and were again cautioned to wait for the "ready" signal and encouraged to push the key as fast as possible on hearing the target sounds. Between each block the subjects were given a two-minute rest period.

Results

The mean reaction times across subjects for all conditions are presented in Table 2. A repeated measure analysis of variance was performed on the data. The only main effect to obtain statistical significance was that of initial phoneme. Subjects responded significantly faster if the item, word or nonword, began with the phoneme /b/ ($\bar{X}$ = 600 msec) than if the item began with the phoneme /s/ ($\bar{X}$ = 736 msec) [$F(1,14) = 16$, $p < .01$]. The word-versus-nonword difference was not significant, although there was a tendency for subjects to respond faster to words ($\bar{X}$ = 683 msec) than to nonwords ($\bar{X}$ = 713 msec) [$F(1,14) = 3.33$, $p < .06$]. There was also a marked, though nonsignificant, tendency for binaural listening to be superior to monaural listening [$F(2,28) = 3.3$, $p < .1$]. The overall error rate was 8.3 percent, but an analysis of variance of the error data revealed no significant difference between the groups. Out of 720 total responses there were 30 errors in both the word and nonword conditions. There was, however, a greater overall percentage of errors in items beginning with /s/ (9.72 percent) than in those beginning with /b/ (6.94 percent).

TABLE 2: Experiment 1: Mean of subjects' mean reaction times in msec.

	Initial phoneme					
	/b/			/s/		
	Ear of presentation			Ear of presentation		
	Left	Right	Binaural	Left	Right	Binaural
Word	648	669	608	744	716	714
Nonword	692	675	668	768	769	[illegible]

The greater difficulty in targeting for /s/ than for /b/ in initial position of a spoken item has been reported previously (Savin and Bever, 1970). A possible source of this difficulty in the present experiment was a reduction of sound quality in the /s/ segments. This was due to the sampling rate of the analog-to-digital converters in the Haskins PCM program. Because of these factors, an analysis of variance independent of /s/ target data was undertaken. This analysis, for only those items with the phoneme /b/ in initial position (see Table 2), revealed that subjects responded significantly faster for words ($\bar{X}$ = 642 msec) than for nonwords ($\bar{X}$ = 678 msec) [$F(1,14) = 4.88$, $p < .05$]. Effects of ear of presentation and the interaction of ear of presentation and word versus nonword did not attain significance.

EXPERIMENT II

Although the word/nonword effect was only marginally significant, the results of Experiment I support the notion that the higher-order properties of a speech event affect the detection of its constituent parts. A second study was designed to further examine this effect using a slightly altered experimental procedure. The change in design represented, in part, an attempt to eliminate the dependence of phoneme targeting on the sequencing of meaningful or nonsense items within a block, that is, on extraneous contextual considerations. Further, the experiment was designed to increase the data base, while also randomizing the predictability of appearance of stimulus items. The last change involved the use of syllables with /s/ in the initial position as foils instead of as target items. This change resulted from the difficulty discussed above.

Method

Subjects. The subjects were nineteen female and eleven male undergraduates at the University of Connecticut. Subjects participated to receive experimental credit in a course on introductory psychology.

Materials and apparatus. The stimuli in this experiment differed from those in Experiment I only in terms of the organization of a block. Each block consisted of 60 items. Within a block there was an equal number of target items with the phoneme /b/ in initial position, foil items with /s/ in initial position, and foil items with various other consonants in initial position. These three types of items were equally divided into words and nonwords. In any word/nonword pair, the form of distinction consisted solely of a change in the final consonant (e.g., /bit/ vs. /bip/). All items were randomly organized throughout a block and each block contained a different order of items. Interstimulus intervals were randomly assigned durations of one, two, three, four, or five seconds. The practice block contained 18 items drawn from the overall stimulus set, organized analogously to an actual test block. The methods of stimulus presentation and of data collection were the same as in the previous experiment.

Procedure. The procedure in Experiment II differed from that of Experiment I in only two ways. First, subjects were informed of the nature of the block organization and the stimulus materials. Second, subjects were instructed to press, as fast as possible, one of two keys when they heard any syllable that started with the phoneme /s/ and to press the other key when they

heard any syllable that began with the phoneme /b/. The subjects were required to target for syllables beginning with /s/ in addition to those beginning with /b/ in order to circumvent the possibility of rapid pressing for any utterance. However, responses for /s/ items were not considered for analysis for the reasons cited above. The keypress procedure consisted of keeping the index finger on a start marker and moving upward and left or upward and right to the appropriate keys. The relation between phoneme and key was couterbalanced across subjects.

Results

Table 3 contains a summary of mean reaction times across subjects for all conditions. In a repeated measure analysis of variance the only main effect to attain significance was that of word versus nonword [$F(1,29) = 69.00$, $p < .001$]. Subjects responded faster to targets in words ($\bar{X} = 593$ msec) than to targets in nonwords ($\bar{X} = 644$ msec). A hint of a similar effect in the detection of final consonants is provided by the work of Steinheiser and Burrows (1973). Once again, there was a tendency for subjects to respond more quickly in the case of binaural presentation. The overall error rate was 2.4 percent. Out of a total of 1800 responses, 22 errors occurred in the nonword condition and 21 in the word condition.

TABLE 3: Experiment 2: Mean of subjects' mean reaction times in msec for /b/ in initial position.

	Ear of presentation		
	Left	Right	Binaural
Word	589	609	580
Nonword	645	655	631

Discussion

As remarked at the outset, perception can be characterized as an orderly progression of mappings from less to more abstract representations. In the case of speech, these representations may be identified as follows: auditory, phonetic, phonological, lexical, syntactic, and semantic (Studdert-Kennedy, 1974). This hierarchy suggests that a response to a particular phoneme could be initiated by the results of processing at the second representational level. But our experiments have shown that the lexical or semantic value of the speech utterance--that is, whether it is a word—affects the latency of phoneme detection. Should we take this to mean that the processing levels we have described are incorrectly arranged? That, contrary to the foregoing account, the lexical and semantic representations, say, precede the phonetic in the temporal course of speech perception, and that the phoneme, therefore, is not a major perceptual unit? Though these conclusions seem anomalous to many students of perception, there are some who would not be especially upset. Both Gibson (1966) and Kolers (1972; Kolers and Perkins, 1975), for example, abhor (for radically different reasons) accounts of perception couched in the language of

atomistic elements and rules. For them, the search for, and arguments about, perceptual units are misguided, as is the treatment of perception as discrete and steplike. In their view, perhaps, our result and the conclusions it suggests about speech perception are not so much anomalous as they are indicative of the inadequacy of the hierarchic, elementaristic formulation of perception.

The necessity for including the above caveat in the present discussion depends in part on the assumption that our experiments actually tap the perception of speech. There is a possibility that this assumption is false; in short, that our experiments do not comment on the nature of perceptual processing at all.

In response to the data obtained from latency-to-detection experiments, Foss and Swinney (1973) drew the following, speculative conclusion: the processes of perceiving are largely separate from and independent of the processes by which perceived events become consciously identified to serve as a basis for differential responding. One of us (Turvey, 1974) has drawn a similar distinction, but in Polanyi's (1964, 1966) terms, that is, between the processes of tacit and explicit knowing. Turvey (1974) conjectures that the processes underlying tacit knowing (perceiving) are different in kind from those underlying explicit identification. From this point of view, the detection task does not reveal perceptual units nor does it tap perceptual processes; to the contrary, it assays operations subsequent to perception. These operations of identification can result in differences in the rates at which descriptions of linguistic events become available to consciousness as a basis for responding (cf. Foss and Swinney, 1973; Ball, Wood, and Smith, 1975).

Consider once again the perception of speech from a hierarchical point of view. There are a number of reasons for proposing that the relations among the levels ought to be quite flexible, with higher-level procedures correcting or verifying descriptions reached tentatively by lower-level procedures (cf. Studdert-Kennedy, 1974). Given the outcome of our experiments, therefore, we can assume from this perspective that there has been considerable confluence among the various levels prior to that identification of the speech event permitting differential responding. But assuming that the representations are at least partially successive and that phoneme detection occurs early on, then why should the response contingent on phoneme detection be delayed until both lower and higher representations have made their contribution? The answer, apparently, is that in the detection task the response mechanism cannot be engaged until perception is complete. In short, even if there do happen to be levels or stages in speech perception, the detection task, unfortunately, will not reveal their order of operation to us. Let us return, therefore, to Foss and Swinney's (1973) thesis.

A cardinal feature of their argument, and one that bears on our particular finding, is that larger language units become available to consciousness--that is, become explicit--sooner than smaller language units (cf. Ball et al., 1975). In the present experiments, all items--both words and nonwords--were monosyllables. If we take the notion of "larger units" literally, then we cannot attribute the latency difference in initial phoneme detection to a hypothesized word/nonword difference in time to access consciousness, since words and nonwords were of the same size. Consequently, if words and nonwords

become available to consciousness at the same latency, then the word advantage effect in initial phoneme detection must be due to a difference in the ease with which words and nonwords as "larger units" can be fractionated into phonemes as "smaller units." However, an alternative interpretation, and one that we prefer, is that a metric of familiarity and/or meaning, rather than one of size, determine the latency of availability to consciousness. In this view, because a word has meaning its description is made available sooner than that of a nonword. The latency difference for initial phoneme detection can then be attributed to this differential rate of availability to awareness of the linguistic description.

REFERENCES

Ball, F., C. Wood, and E. E. Smith. (1975) When are semantic targets detected faster than visual or acoustic ones? Percept. Psychophys. 7, 1-8.

Biederman, I. (1972) Perceiving real-world scenes. Science 177, 77-79.

Biederman, I., A. L. Glass, and E. W. Stacy, Jr. (1973) Searching for objects in real-world scenes. J. Exp. Psychol. 97, 22-27.

Cooper, F. S. and I. G. Mattingly. (1969) A computer controlled PCM system for the investigation of dichotic perception. J. Acoust. Soc. Am. 46, 115(A).

Foss, D. J. and D. A. Swinney. (1973) On the psychological reality of the phoneme: Perception, identification, and consciousness. J. Verbal Learn. Verbal Behav. 12, 246-257.

Gibson, J. J. (1966) The Senses Considered as a Perceptual System. (Boston: Houghton Mifflin).

Hadding-Koch, K. and M. Studdert-Kennedy. (1964) An experimental study of some intonation contours. Phonetica 11, 175-185.

Kolers, P. A. (1972) Aspects of Motion Perception. (New York: Pergamon Press).

Kolers, P. A. and D. N. Perkins. (1975) Spatial and ordinal components of form perception and literacy. Cog. Psychol. 7, 228-267.

McNeill, D. and K. Lindig. (1973) The perceptual reality of phonemes, syllables, words and sentences. J. Verbal Learn. Verbal Behav. 12, 419-430.

Polanyi, M. (1964) Personal Knowledge: Towards a Post-Critical Philosophy. (New York: Harper & Row).

Polanyi, M. (1966) The Tacit Dimension. (Garden City, N. Y.: Doubleday).

Reicher, G. M. (1968) Perceptual recognition as a function of meaningfulness of stimulus material. Technical Report No. 7. (The University of Michigan, Human Performance Center).

Savin, H. B. and T. G. Bever. (1970) The nonperceptual reality of the phoneme. J. Verbal Learn. Verbal Behav. 9, 295-302.

Steinheiser, F. H., Jr. and D. J. Burrows. (1973) Chronometric analysis of speech perception. Percept. Psychophys. 13, 426-430.

Studdert-Kennedy, M. (1974) Speech perception. Haskins Laboratories Status Report on Speech Research SR-39/40, 1-52.

Studdert-Kennedy, M. and K. Hadding. (1973) Auditory and linguistic processes in the perception of intonation contours. Lang. Speech 16, 293-313.

Turvey, M. T. (1974) Constructive theory, perceptual systems, and tacit knowledge. In Cognition and the Symbolic Processes, ed. by W. B. Weimer and D. S. Palermo. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.), pp. 165-180.

Weisstein, N. and C. S. Harris. (1974) Visual detection of line segments: An object-superiority effect. Science 186, 752-754.

Wheeler, D. D. (1970) Processes in word recognition. Cog. Psychol. 1, 59-85.

On Detecting Nasals in Continuous Speech*

Paul Mermelstein

ABSTRACT

The acoustic manifestation of nasal murmurs is significantly context dependent. To what extent can the class of nasals be automatically detected without prior detailed knowledge of the segmental context? This contribution reports on the characterization of the spectral change accompanying the transition between vowel and nasal for the purpose of automatic detection of nasal murmurs. The speech is first segmented into syllable-sized units, the voiced sonorant region within the syllable is delimited, and the points of maximal spectral change on either side of the syllabic peak are hypothesized to be potential nasal transitions. Four simply extractible acoustic parameters, the relative energy change in the frequency bands 0-1, 1-2, and 2-5 kHz, and the frequency centroid of the 0-500-Hz band, at four points in time spaced 12.8 msec apart, are used to represent the dynamic transition. Categorization of the transitions using multivariate statistics on some 524 transition segments from data of two speakers resulted in a 91 percent correct nasal/nonnasal decision rate.

INTRODUCTION

The search for invariant acoustic cues that indicate the presence of nasal murmurs in continuous speech has a long history. Fujimura (1962) reported the spectral characteristics of nasal murmurs in intervocalic contexts. He found three essential features: first, the existence of a very low first formant in the neighborhood of 300 Hz; second, the relatively high damping factors of the formants, and third, the high density of the formants in the frequency domain. Fant (1962) reports that a voiced occlusive nasal (nasal murmur) is characterized by a spectrum in which the second formant is weak or absent; a formant at approximately 250 Hz dominates the spectrum, but several weaker high-frequency formants occur, and the bandwidths of nasal formants are generally larger than in vowel-like sounds. These cues appear to be generally adequate for human identification of nasal segments in spectrograms; however, precise quantitative data are unavailable. This work represents an attempt to evaluate the contributions of a set of simply extractible acoustic measurements to the detection of nasals.

*A version of this paper was presented at the 90th annual meeting of the Acoustical Society of America, San Francisco, Calif., 3-7 November 1975. [J. Acoust. Soc. Am. (1975), Suppl. 58, S97(A).]

Acknowledgment: This research was supported in part by the Advanced Projects Agency of the Department of Defense under contract No. N00014-67-A-029-002 monitored by the Office of Naval Research.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

We approach the problem as a typical categorization task in automatic phonetic transcription. A general framework for the solution of this task has been presented previously (Mermelstein, 1975a). The point of view adopted there emphasizes the hierarchical nature of speech perception and structures the automatic phonetic analysis tasks analogously. Detection of nasals is one stage of that process. The continuous speech signal is first segmented into syllable-sized segments called syllabic units. The syllabic unit may be entirely nasal, a syllabic nasal. Otherwise, constraints on the phonetic structure of syllables allow the existence of at most one manner of production change from nasal to nonnasal prior to the syllabic peak and one reverse change from nonnasal to nasal after the syllabic peak. We first delimit the voiced sonorant portion of the syllable and look for points of maximal spectral change within the delimited interval on either side of the syllabic peak. These points of maximal spectral change are hypothesized as potential transition points between the vowel (possibly also glide or liquid) and the nasal. The detected transitions are categorized on the basis of acoustic measurements in the transition region.

What are the acoustic cues that allow the listener to establish that a particular syllable contains a nasal segment? A preliminary exploratory study used bisyllabic nonsense words with nasals in intervocalic environment as well as in intervocalic clusters where the nasal preceded or followed a stop consonant. Examination of spectrograms and spectral cross sections essentially confirmed Fujimura's (1962) report. Reliable cues for nasals are found to be a low-frequency nasal resonance and a drop in mid- + high-frequency energy (above roughly 1000 Hz) in the absence of a significant drop in low-frequency energy (below 1000 Hz). Suitable qualitative parameter differences were easily found by inspection. However, when the same cues were tested on continuous speech, differentiation between nasals and nonnasals proved remarkedly poorer. Accordingly, a new study was carried out in an attempt to characterize quantitatively these parameters in continuous speech and evaluate their utility for nasal detection.

The distinctive manner feature "nasalized" pertains to both nasal vowels and nasal murmurs. This study is concerned with the transition from vowel (nasalized or not), glide or liquid, to nasal murmur, where the primary articulatory change is oral closure in the absence of velopharyngeal closure. Instead of searching separately for the acoustic correlates of the oral closure and velopharyngeal opening, which can be expected to show gross variations depending on the state of the other features, it appears worthwhile to look for correlates of the composite articulatory event directly.

The actual spectrum of the nasal murmur is known to vary with the syllabic vowel as well as the place of oral closure (Fujimura, 1962). Since place-of-production discrimination can be expected to be highly manner-of-production dependent, we chose first to detect nasal murmurs as a class and subsequently to discriminate among them.

The first parameter selected, the energy centroid in the 0-500-Hz frequency band, can be looked on as a rough approximation to the first-formant frequency. A value for this parameter near 250 Hz is a necessary but not sufficient condition for the existence of nasal murmurs because this property is shared by the first-formant frequency of high vowels. The energy parameters defined below are intended to discriminate between the nasals and the high vowels. The energy centroid, although independent of overall signal level, is dependent on linear spectral distortion such as the 300-Hz high-pass filtering of telephone speech.

Fant (1967) suggests that the physical phenomena underlying a particular distinctive feature need exhibit only relational invariance. For example, the weakness of a second formant may be best judged relative to the intensity of that formant in the adjacent vowel rather than in absolute terms. We employ three spectral energy parameters, all defined in relational terms with respect to the energy in the respective frequency bands prior to the transition. This definition makes the parameters independent not only of the overall signal amplitude as well as any linear spectral distortion, but corrects to a limited extent for the overall spectral shape imposed by the syllabic vowel. Since none of the parameters alone is sufficiently effective to separate the nasals, our effort has focused on the effective combination of information from several independently measured parameters in an attempt to attain classification performance superior to that obtainable by any single parameter.

EXPERIMENTAL PROCEDURE

In order to examine nasals in a wide variety of vocalic and consonantal contexts, previous recordings of the "rainbow passage" and six additional sentences by two speakers were studied. The speech material was recorded at the subjects' comfortable reading rate, digitized using a 10-kHz sampling frequency, and spectra were computed using a 25.6-msec Hamming time window. Adjacent spectral computations were spaced 12.8 msec apart and yielded results with 40-Hz frequency spacing. The material was segmented into syllabic units following procedures reported separately (Mermelstein, 1975b).

To test the hypothesis that points of maximal spectral change are potential nasal indicators, we need to define operationally the term "syllabic peak" and an appropriate metric for "spectral derivative." It is our intent that the syllabic peak be located within the vocalic region of any syllable at the point of minimal spectral change so that it best reflects the color of the syllabic vowel. Having established this point of minimal spectral change within the vocalic region, the spectral derivative within the voiced regions on either side of the syllabic peak can be computed and maxima found. By evaluating the acoustic information in the neighborhood of the maximal spectral changes, we shall try to classify the transition as to whether it denotes the onset or termination of a nasal.

The syllabification program evaluates minima in a "loudness function" (a time-smoothed, frequency-weighted energy function) as potential syllabic boundaries. The maxima in loudness are potential syllabic peaks. Qualitative study of spectrograms augmented with loudness curves reveals that frequently the maximum in loudness occurs prior to the time at which the formants appear to be maximally steady. Hence we construct a 6-dB loudness range below the maximal loudness level of the syllabic unit and search for the point of minimal spectral change within the corresponding time interval. Figure 1 shows typical plots of loudness and spectral differences for one segment. The definition of spectral variation with time is guided by the following rationale. Pols (1972) has computed the eigenvectors accounting for the two dimensions of maximal variance for sonorants. These roughly correspond to measures of speech spectra along the dimensions of low-frequency versus high-frequency energy and the low- and high-frequency versus midfrequency difference. In an attempt to approximate perceptually equal changes in vowel spectra by equal increments in our two-dimensional spectrum representation, we first transform the sepctra from a linear frequency scale to a technical mel scale (linear up to 1000 Hz, logarithmic thereafter).

Next we compute the first two coefficients of the Fourier cosine transform of the log power spectrum of the signal using the weighting functions shown in Figure 2. The directions in multidimensional spectral space defined by these coefficients roughly correspond to the maximal variance directions of Pols (1972). Our dimensions are orthogonal and coefficient values are independent of spectrally uniform signal amplification or attenuation. Individual spectra at time k can now be represented by the coefficient pair

$$C_i(k) = \int_{f=0}^{f=4\ \text{kHz}} w_i(f)\, e_k(f), \quad i = 1, 2$$

where the w_i are the respective weighting functions and $e_k(f)$ is the measured energy as a function of frequency at time k.[1]

Let k_a and k_b be the boundaries of the voiced, nonfricative central section of a syllabic unit. Now define the spectral difference metric at time-sample k as

$$D(k) = \left[\sum_{i=1}^{2} \left(C_i(k+1) - C_i(k-1)\right)^2\right]^{1/2}$$

where the $C_i(k)$ are the coefficients computed for the k^{th} spectral cross sections and unit spacing in k corresponds to a time spacing of 12.8 msec. This is the same metric as that used by Itahashi, Makino, and Kodo (1973) for phonetic segmentation, except that our definition is symmetric with time so that it is independent of movement forward or backward in time from the syllabic peak.[2]

Define the concave hull of the difference function over the interval k_a-k_p-k_b by

---

[1] If the logarithmic frequency-scale transformation were omitted, the computed coefficients would correspond to the first and second coefficients in a real cepstrum (cosine transformation of power spectrum) representation of the speech signal. The zeroeth coefficient corresponds to the average spectrum level and is therefore not used. Truncation of the cepstrum at a point in inverse-time (quefrency) lower than the pitch period yields a smoothed spectrum envelope. The first two coefficients capture the most significant aspects of the variations of spectrum envelope with frequency. Preliminary evaluation shows separation of vowels in the two-coefficient space comparable to a two-formant representation at a significantly reduced computational cost.

[2] Since the C_i coefficients are linear functions of the signal energy, the difference metric could be equally well-defined in terms of weighted spectral differences. The C_i coefficients are initially computed for the purpose of syllabic-vowel categorization, a task not discussed in this report. The computation of the difference metric D(k) is but a simple additional step.

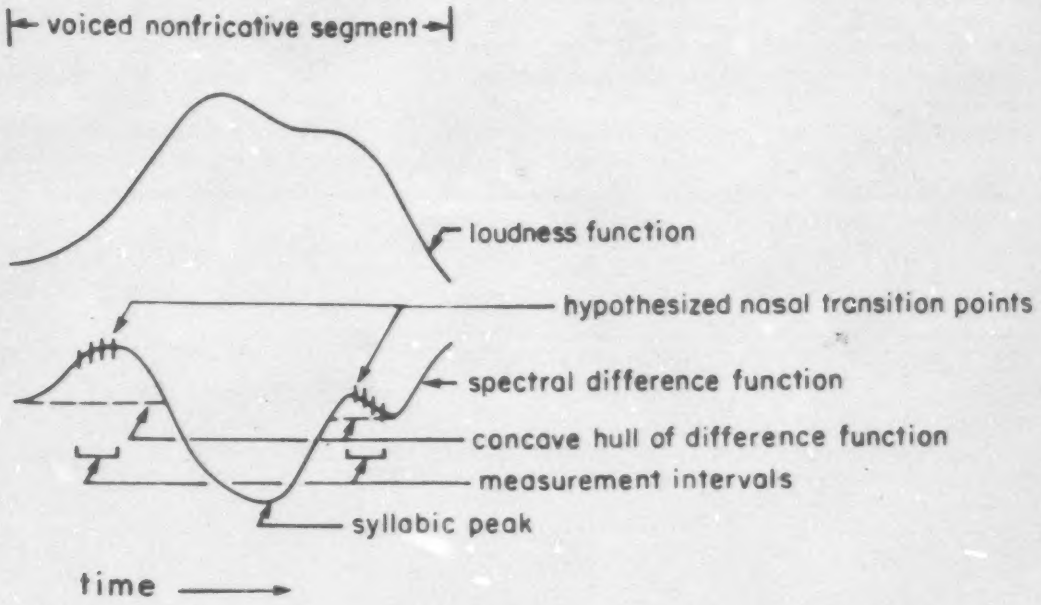


Figure 1: Loudness and spectral difference functions for a typical segment.

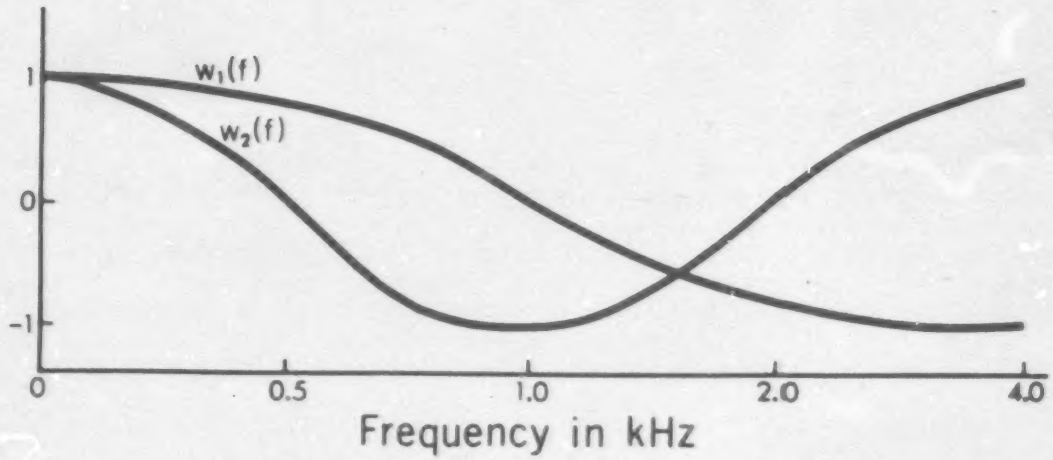


Figure 2: Weighting functions for the determination of spectral coefficients.

$$H(k) = \min_{k_a \le k' \le k} D(k'), \quad k_a \le k < k_p$$

$$= \min_{k < k' \le k_b} D(k'), \quad k_p \le k \le k_b$$

where k_p is the time frame of the syllabic peak. Now find the maximum difference in the spectral difference less the null, on either side of the syllabic peak

$$D'(k_r') = \max_{k_a \le k < k_p} [D(k) - H(k)]$$

$$D'(k_q') = \max_{k_p \le k \le k_b} [D(k) - H(k)]$$

Then $k_q = k_q' - 1$ and $k_r = k_r' + 1$ are points of potential nasal onset or termination. Now consider the acoustic characteristics of the spectral regions on the time-positive side of k_q and the time-negative side of k_r. For notational convenience, let α be -1 near k_r, +1 near k_q, and consider the sepctral frames $S_n(k_q + \alpha n)$, $S_n(k_r + \alpha n)$, $n = 1, \ldots, 4$. If there is a nasal segment in time-final position in the syllabic unit, then the spectra S_n can be expected to reflect that manner category.

We now define our basic measurements. Let $\Delta E_n^i = E_{k+n}^i - E_k^i$ be the relative energy (dB) in the i^{th} frequency band of the n^{th} frame relative to the energy in the same frequency band at the onset or termination of the hypothesized segment. Approximate the first formant at time k + n by the centroid of the energy in the frequency band 0-500 Hz,

$$g_n = \sum_i f_i \, e_{k+n}(f_i) / \sum_i e_{k+n}(f_i)$$

Our analysis system outputs filter spectra at 40-Hz intervals, thus the summation over i ranges over the first 12 spectral samples. The above measurements can be seen to be simply derivable from the speech signal even under relatively noisy conditions and were therefore considered to be potentially robust cues for nasal segments. The question to be investigated is to what extent the parameters ΔE_n^i and g_n differentiate the nasals from the nonnasals, and thus represent useful cues for automatic nasal detection.

The information contributed by the various parameters at the respective points in time may be combined in diverse ways. The simple statistical model used for our initial attempts at classification treated the cues as independent time-varying quantities. The relative likelihood that the transition belongs to the nasal or nonnasal class is computed by summing the relative likelihood scores arrived at from each parameter at each point in time. However, the parameters are in fact correlated. For example, transitions to obstruents are accompanied by a large energy drop and a low low-frequency centroid. Furthermore, liquids show a significant drop in the 2-5-kHz band, but much less of a drop in the 1-2-kHz band. An improved statistical model uses multivariate statistics on all parameters at the respective points in time. Since the time course of parameter

variation may be different for the various parameters, separate likelihood scores are computed at each point in time and summed to result in a composite score.

In the absence of a priori information regarding the distribution of the acoustic parameters, we assumed multivariate normal distributions and estimated the parameter mean vectors $\underline{m}_a^n$, $\underline{m}_b^n$ and the parameter covariance matrices $\underline{\Sigma}_a^n$ and $\underline{\Sigma}_b^n$ for the nasal and nonnasal transitions, respectively. The superscript n denoted the point in time for the parameter measurement. Data were collected for n = 1,..., 4, a time frame of 51 msec of spectral data, which in turn is derived from some 64 msec of waveform data.

Following Patrick (1972), the minimum probability of error decision rule for two categories when each has a Gaussian distribution with estimated mean and covariance matrix is to decide Class "a" if

$$\frac{P_a}{\left[\underline{\Sigma}_a\right]^{1/2} (2\pi)^{L/2}} \exp\left[-\frac{1}{2}(\underline{x} - \underline{m}_a)^t \underline{\Sigma}_a^{-1}(\underline{x} - \underline{m}_a)\right] >$$

$$\frac{P_b}{\left[\underline{\Sigma}_a\right]^{1/2} (2\pi)^{L/2}} \exp\left[-\frac{1}{2}(\underline{x} - \underline{m}_b)^t \underline{\Sigma}_b^{-1}(\underline{x} - \underline{m}_b)\right]$$

where P_a and P_b are the a priori probabilities of the respective categories, and L is the dimensionality of the measurement vector $\underline{x}$. To improve the reliability of our decisions as well as to make them insensitive to small registration errors in the time signal, we wish to combine information from parameter measurements at several points closely spaced in time. Since we are dealing with a dynamic articulatory event, parameter statistics must be assumed nonstationary. Instead of treating our parameters as multivariate in space and time--an operation that would require significantly more parameter estimation data than available--independent estimates of parameter means and covariances were carried out at each measurement time. One decision rule that combines this information (decision rule A) is the following: if

$$P_a \prod_n \left[\underline{\Sigma}_a^n\right]^{-1/2} \exp\left[-\frac{1}{2}(\underline{x}^n - \underline{m}_a^n)^t \left(\underline{\Sigma}_a^n\right)^{-1} (\underline{x}^n - \underline{m}_a^n)\right] >$$

$$P_b \prod_n \left[\underline{\Sigma}_a^n\right]^{-1/2} \exp\left[-\frac{1}{2}(\underline{x}^n - \underline{m}_b^n)^t \left(\underline{\Sigma}_a^n\right)^{-1} (\underline{x}^n - \underline{m}_b^n)\right]$$

choose category a, otherwise choose category b. Here the superscript n denotes the measurement time.

Alternatively, we may wish to normalize the relative probabilities before summation over the different measurements. For this decision rule (decision rule B), define a category score

$$s^n(\alpha) = \frac{p^n(x|\alpha)}{p^n(x|\alpha) + p^n(x|\beta)}, \quad 0 \leq s(\alpha) \leq 1; \ \alpha = a,b; \ \beta = b,a$$

and if $P_a \sum_n s^n(a) > P_b \sum_n s^n(b)$, choose a, otherwise choose b.

The effects of the a priori probabilities P_a and P_b may be embedded in a decision threshold θ and adjustment of θ up or down may be used to control the difference between the relative frequency of false nasal and nonnasal decisions. To obtain the results cited, a value of θ was used that results in roughly equal probability of false nasal and nonnasal decisions.

RESULTS

A preliminary analysis program found the points of maximal spectral difference. On the basis of spectrographic and auditory examination, these transition points were hand-labeled to indicate whether they corresponded to nasal or nonnasal transitions. A single nasal segment could be manifested by two transitions if in intervocalic context, and one transition only if in a pre- or post-obstruent context. Syllabic units were treated as independent information-bearing elements and each transition was classified independently. Two syllabic nasals were found in the data and these were eliminated from subsequent consideration.

Statistics were gathered separately for nasal-nonnasal and nonnasal-nasal transitions. Differences between nasals in initial and final position in the voiced sequences of the syllabic unit were not found significant, and the two classes were therefore pooled to arrive at the following results. Figure 3 gives means and standard deviation values for the measured parameters after pooling of the differently directed transition groups. One observes that the distributions of all of the parameters show considerable overlap. Only for ΔE_n^1 do we see considerable separation by categories. However, ΔE_n^1 does not separate the nonnasal sonorants from the nasals. It only serves to exclude the transitions to nonsonorants. Parameter g_n shows little separation in category means but a large difference between the variances of the two categories. In fact, detailed examination shows the distribution of the nonnasals to be roughly bimodal; the obstruents have rather low values of g_n, the sonorants have values higher than the mean nasal value. Clearly, the nonnasal category is not homogeneous and perhaps a representation of the nonnasals in terms of a mixture of normally distributed categories would be more appropriate.

The decision threshold for the two categories was established by pooling the parameter data of both subjects and using the same 524 transitions both to train the classifier and to test it. A misclassification rate of 13.9 percent was observed for decision rule A. The procedure was repeated with decision rule B and resulted in 9.3 percent errors. Evidently, normalization of the conditional probabilities before combining the measurements at different points in time helps to lower the error rate.

Experiments were continued with decision rule B. Speaker to speaker variation was estimated by repeatedly testing one speaker's data against measurements derived from the other speaker's data. A total error rate of 15 percent was noted. The most significant differences in the nasal transition parameter data

FIGURE 3

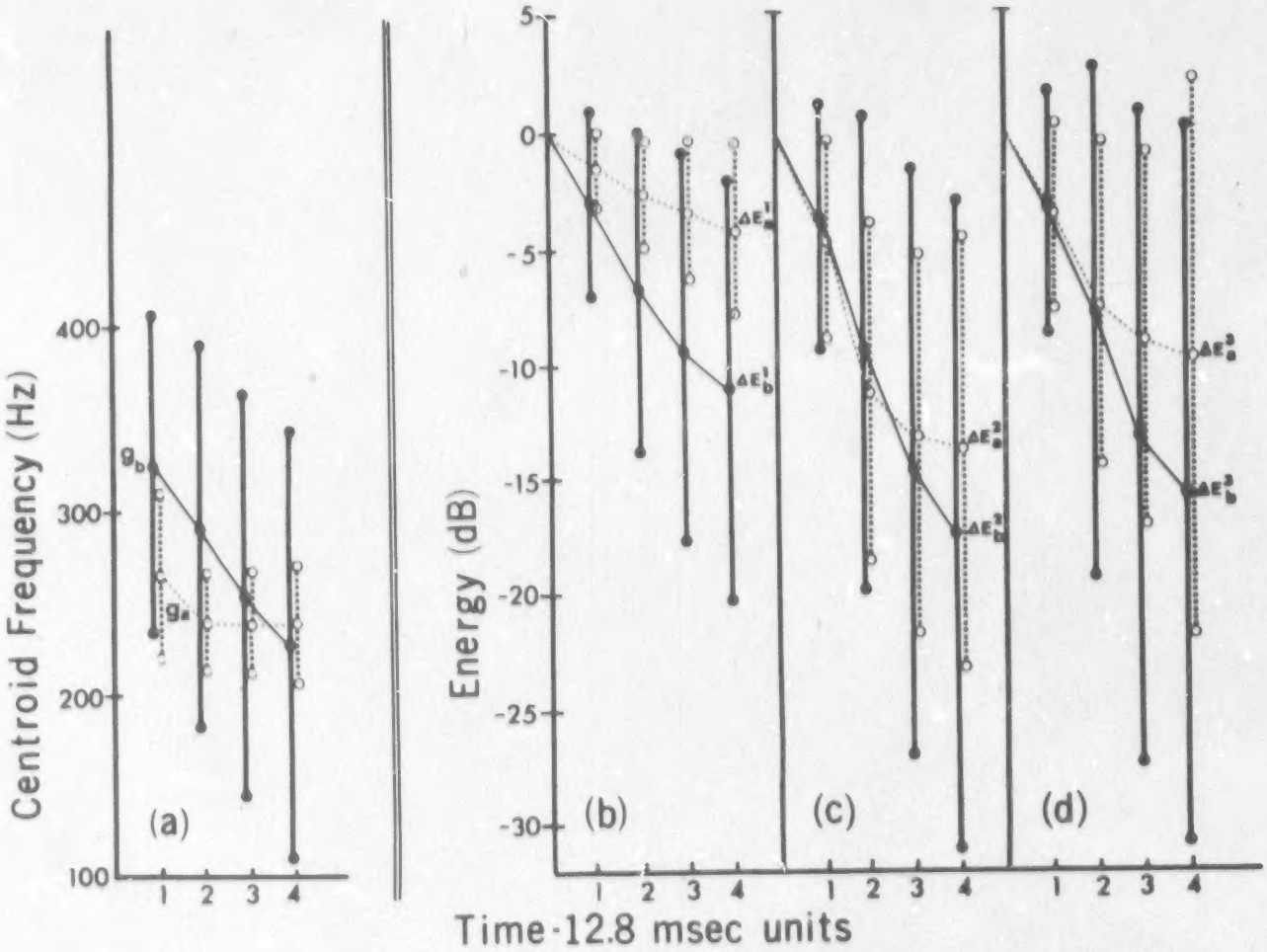


Figure 3: Mean and standard deviation values for measured parameters. (a) Centroid frequency, 0-500-Hz band; (b) relative signal energy, 0-1-kHz band; (c) relative signal energy, 1-2-kHz band; (d) relative signal energy, 2-5-kHz band. Subscript a - nasal category, b - nonnasal category.

were noted in the mean centroid frequency. For n = 2, this value was 222 ± 16 Hz for speaker LL, 256 ± 26 Hz for speaker GK. As discussed below, this parameter was found to be the most useful contributor to the total categorization score. Thus it is not surprising that the decisions are strongly dependent on small centroid frequency differences. Based on articulatory considerations, one would expect significant nasal resonant-frequency differences due to size differences between speakers' nasal cavities. The relatively small standard deviation values for the parameter are more surprising and indicate the insensitivity of this parameter to contextual variations. When training data and test data were separated by text material rather than speaker, the total error rate was only 11 percent. The higher error-rate degradation due to learning and testing on different speakers rather than different text suggests that further improvements in categorization may result through use of speaker-dependent measurement data.

To evaluate the relative contributions of the four measurements, a decision rule was implemented that treated each measurement as independent, normalized the conditional probabilities for each measurement, and summed to contributions from the 16 measurements. Predictably the error rate on the total data using independent parameters was higher: 13.5 vs. 9.3 percent, using multivariate statistics. An estimate of the contribution of each measurement to the total decision score may be derived from

$$S^i = \frac{1}{J} \sum_{j=1}^{J} \beta_j \, [s_j^i(a) - s_j^i(b)]$$

where β_j is +1 or -1 depending on whether or not the test item was a nasal, $s_j^i(a)$ and $s_j^i(b)$ are the respective normalized probability scores for the two categories obtained from measurement i on token j, and J is the total number of transition tokens. The low-frequency energy centroid--the one parameter dependent on the spectrum at only one moment--showed the highest contributions, namely, 0.71, 0.77, 0.75, and 0.74 for n = 1, ..., 4. The other parameters were apparently less effective; their contributions ranged from 0.53 to 0.65. Measurements at time values n = 2, 3 were most effective, yet the others still contributed substantially to reduce the overall error rate. The one significant difference between prevocalic and postvocalic nasal transition was found in the relative effectiveness of the measurements at the distinct time values. Measurements at small n values give relatively higher contributions for prevocalic transitions, measurements at larger n values are more effective for postvocalic transitions. One explanation for this may be that a nasal is frequently anticipated by nasalization of the preceding vowel, which causes the spectral discontinuity to be less abrupt and requires a longer time delay before a distinct nasal murmur can be observed.

DISCUSSION

In attempting to compare our results with these of other workers, we encountered few quantitative results in the literature. Weinstein, McCandless, Mondsheim, and Zue (1975) report confusion statistics for consonant segment classification. If their confusion matrix is reduced to two categories--nasal and nonnasal--a misclassification rate of 21 percent is obtained. The test applied there to detect nasals makes use of information similar to that in our work. However, two essential differences should be noted. First, by averaging formant frequency and amplitude measurements over five points in time before

classification, they implicitly assume a nasal-segment model with static spectral characteristics. Our data reveal significant parameter differences with time as one moves further into the nasal segment. Second, formant computations appear not to be necessary for nasal detection. In fact, representation of the nasal spectrum by means of three formants may not be sufficiently precise. One of the differentiating characteristics of nasals is the presence of spectral zeros, the effects of which are poorly captured in a three-formant representation. In view of our results, the use of broad-band spectral information appears more robust.

Generalization of the nasal/nonnasal discrimination to further speakers must await the collection and processing of further data. Interspeaker variation appears to be the most significant limitation to improvement of the classification results. We suspect that a limited amount of unsupervised training may suffice to overcome this limitation; however, no experimental studies of this question have been carried out.

There are two important additional sources of variance in our data. The nasal spectrum depends on the color of the syllabic vowel because that is the underlying articulation on which the nasal murmur articulation is superimposed. Of course, the nasal spectrum further depends on the place of production of the nasal. No attempts to use our measurements to categorize the nasal murmurs by place of production have yet been carried out. Because good nasal/nonnasal classification is obtainable without consideration of place or production information, it appears appropriate for any complete analysis to do nasal/nonnasal classification first, followed by categorization of the nasal segments.

Most of the false indications result from the confusion of liquids, glides, and semivowels with nasals. In particular, /l/ and /r/ before high vowels tend to be confused with nasals rather often. In addition, some voiced fricatives that manifest weak frication, particularly in unstressed environments, can be confused with nasals. Nasals were missed most frequently when they appeared to be shortened owing to a consonantal cluster context or when they appeared to be articulated as a nasal flap. In cases where nasals are shorter than 50 msec, summation of partial scores from four points in time may be inferior to a sequential classification procedure that stops consideration of new measurements whenever the partial sum of scores exceeds a given fraction of the total possible score.

CONCLUSIONS

The spectral changes manifested by the transitions to and from nasal murmurs are good cues for the recognition of the nasals as a class. Of the four measurements used, the centroid in the 0-500-Hz frequency band appears to be the most useful parameter. Use of additional measurements of energy change in three broad frequency bands allows good separation of nasals and nonnasals irrespective of context. The measurements are significantly correlated, thus resort to multivariate statistics is necessary.

It appears particularly important to treat the transition between nasal and nonnasal as a dynamic articulatory event with corresponding time-varying acoustic properties. The individual parameters show significant variation with increasing time displacement from the onset of the transition.

Maximal separation between nasals and nonnasals is not achieved at the same point in time for all the parameters. Therefore, the data must not be pooled over the separate time points of measurement. Through careful selection of the maximal spectral variation point, we achieve a time synchronization of the unknown transition with respect to the corresponding reference data and thereby obtain improved separation between the nasal and nonnasal categories.

REFERENCES

Fant, G. (1962) Descriptive analysis of the acoustic apsects of speech. Logos 5, 3-17.

Fant, G. (1967) The nature of distinctive features. In For Roman Jakobson. (The Hague: Mouton).

Fujimura, O. (1962) Analysis of nasal consonants. J. Acoust. Soc. Am. 34, 1865-1875.

Itahashi, S., S. Makino, and K. Kodo. (1973) Discrete-word recognition utilizing a word dictionary and phonological rules. IEEE Trans. Audio Electroacoust. AU-21, 239-249.

Mermelstein, P. (1975a) A phonetic-context controlled strategy for segmentation and phonetic labeling of speech. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-23, 79-82.

Mermelstein, P. (1975b) Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. 58, 880-883. [Also in Haskins Laboratories Status Report on Speech Research SR-42/43 (1975), 247-256.]

Patrick, E. A. (1972) Fundamentals of Pattern Recognition. (Englewood Cliffs, N. J.: Prentice-Hall).

Pols, L. C. W. (1972) Segmentation and recognition of mono-syllabic words. Conference Record, Conference on Speech Communication and Processing, IEEE and Air Force, Cambridge Research Laboratories, pp. 105-108.

Weinstein, C. J., S. S. McCandless, L. F. Mondsheim, and V. W. Zue. (1975) A system for acoustic-phonetic analysis of continuous speech. IEEE Trans. Acoust. Sig. Proc. ASSP-23, 54-67.

A Digital Pattern Playback for the Analysis and Manipulation of Speech Signals

P. W. Nye, L. J. Reiss, F. S. Cooper, R. M. McGuire, P. Mermelstein, and T. Montlick

ABSTRACT

The Digital Pattern Playback (DPP) is a new research tool for the analysis, manipulation, and resynthesis of speech data. Based on an interconnected PDP-11/45 and GT40 computer system, the device provides most of the best features of the original pattern playback in conjunction with ready access to a basic collection of signal processing algorithms and input-output facilities. The user works with gray-scale digital displays of spectrographic data generated by a specially built display processor. Conversion of the spectrogram back into an acoustic signal is achieved by means of a channel vocoder. From its inception, a major design objective of the DPP has been to achieve a very short delay between making a change in the speech data and both seeing and hearing the result.

INTRODUCTION

The complex multidimensional structure of the natural speech signal has been a research interest of phoneticians and communications engineers for many years. Among their major concerns is the problem of identifying its essential information-bearing elements--sometimes called the speech cues. The underlying reasons for this work have been threefold: first, it provides a basis for understanding the perception of speech; second, it is a logical step toward the detection of phones and phonemes for continuous speech recognition; and third, knowledge of these cues is essential to the development of speech synthesizers and synthetic speech. Frequently, these researchers found that the search for the natural cues requires a more detailed examination of the speech signal than is possible by ear, and this has led to the development of several specialized analysis instruments, notable among which is the sound spectrograph. In addition, to satisfy the need to verify candidate acoustic features, several synthesis instruments, such as the pattern playback (Cooper, 1953), have been developed. These permit the generation of artificial speech sounds or stimuli

---

Acknowledgment: Work on the original design of the DPP and a portion of its development was supported by grant No. GS-28354 from the National Science Foundation. The authors take pleasure in acknowledging the many original contributions made by E. Wiley and D. Zeichner to the design of the DPP hardware and thank R. Sharkany for his work on its construction.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

containing variants of the features being studied. The stimuli are then subjected to perceptual analysis by groups of listeners whose responses provide a measure of the characteristics or importance of the features.

An important element in the application of these analysis techniques has been not only the availability of appropriate tools but also their operating speed and flexibility. Thus, the most successful facilities have made complete cycles around the analysis-synthesis-analysis loop possible in times that are conveniently short for the experimenter. Moreover, they have provided good visual displays of the modifications being applied to speech cues and have kept the experimenter in close touch with the processes being performed by the analysis hardware.

This paper describes the construction and operation of a special-purpose speech analysis and synthesis tool that has been built for speech research at Haskins Laboratories. Called the Digital Pattern Playback (DPP), the device adheres closely to the philosophy that underlays the original pattern playback and retains most of its essential features, although it is implemented with modern technology. The DPP consists of a combination of hardware and software modules built around a Digital Equipment Corporation (DEC) PDP-11/45 and GT40 interconnected computer system. To achieve acceptably fast operating characteristics, hardware devices are employed for those operations where speed can be used to advantage or is required for an acceptable dynamic display. Software is used on all other tasks where computational accuracy and flexibility are needed. The DPP has been specifically designed to be used in two types of research activity: in the search for natural speech cues associated with basic research on speech perception and in connection with the development of automatic feature extraction algorithms that can be employed at the first stage of a speech recognition system. Examples of both of these applications will be described in more detail later in this paper. This description of the DPP will be approached in three stages.

Stage 1 outlines the environment of the DPP by introducing the key components of the PDP-11/45 and GT40 computer system around which the DPP is built. In stage 2, the structure and function of the specially built hardware components of the DPP are described and, finally, in stage 3 the organization and operation of the DPP software are outlined.

STAGE 1: THE DPP ENVIRONMENT

Figure 1 provides an overall view of the DPP as the user might conceive it, with arrows indicating the direction in which information flows between the major components. A more detailed system plan of the DPP's computer environment is given in Figure 2. A PDP-11/45 with a floating point processor, four disk drives, 96K of core memory, and a complement of other peripherals is connected by two communication pathways to a GT40 display system. The GT40 display terminal consists of a PDP-11/10 Central Processing Unit (CPU), 8K of core memory, a cathode-ray display tube, a light-pen, a display controller, and a keyboard. Connected directly to the GT40 are 16K of auxiliary core memory (memory A) and, via a bus repeater, an extension of the unibus that permits the attachment of additional peripheral components. Among the most important of these components are a hardware spectrum analyzer (UBIQUITOUS manufactured by Federal Scientific), a magnetic writing tablet (manufactured by Summagraphics),

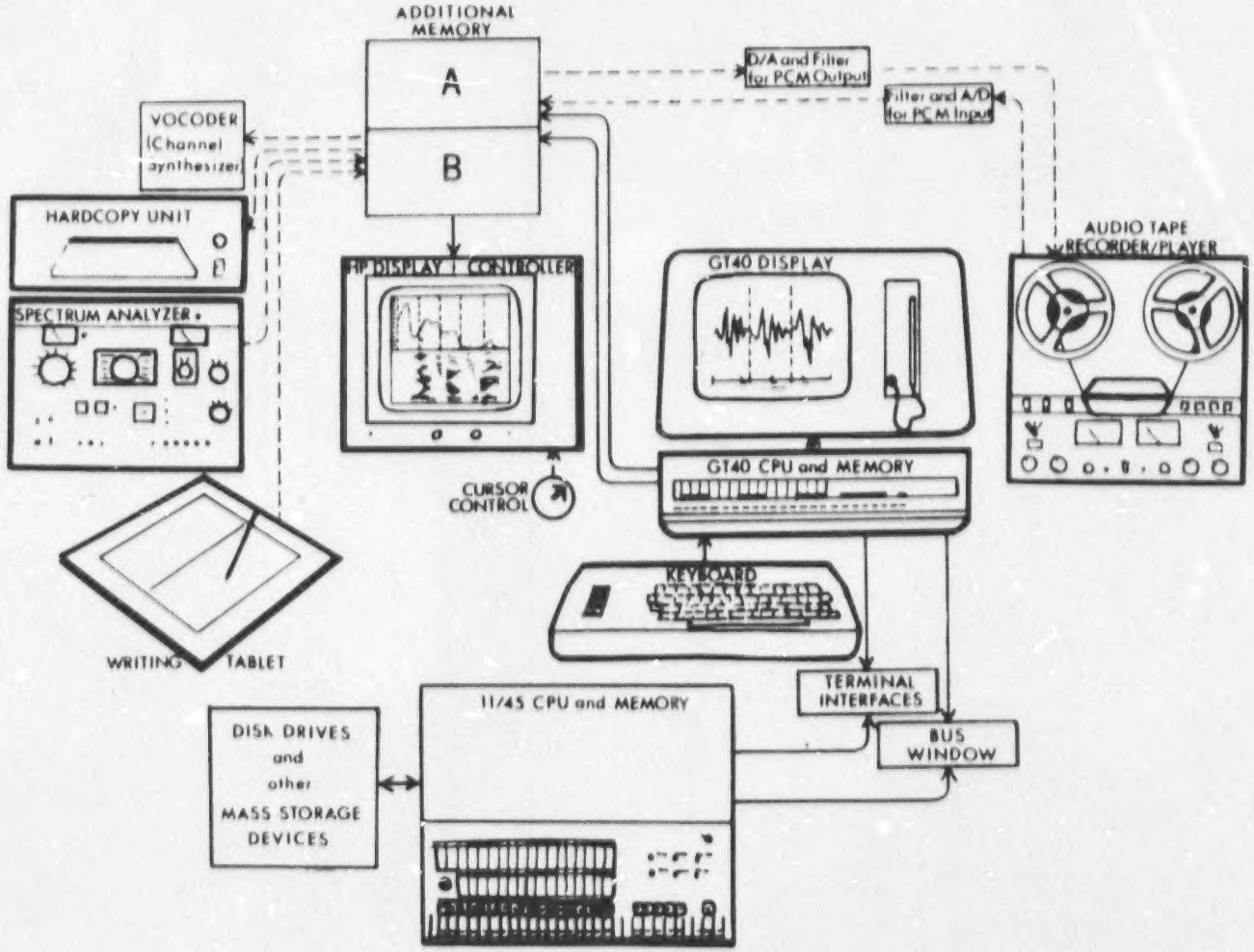
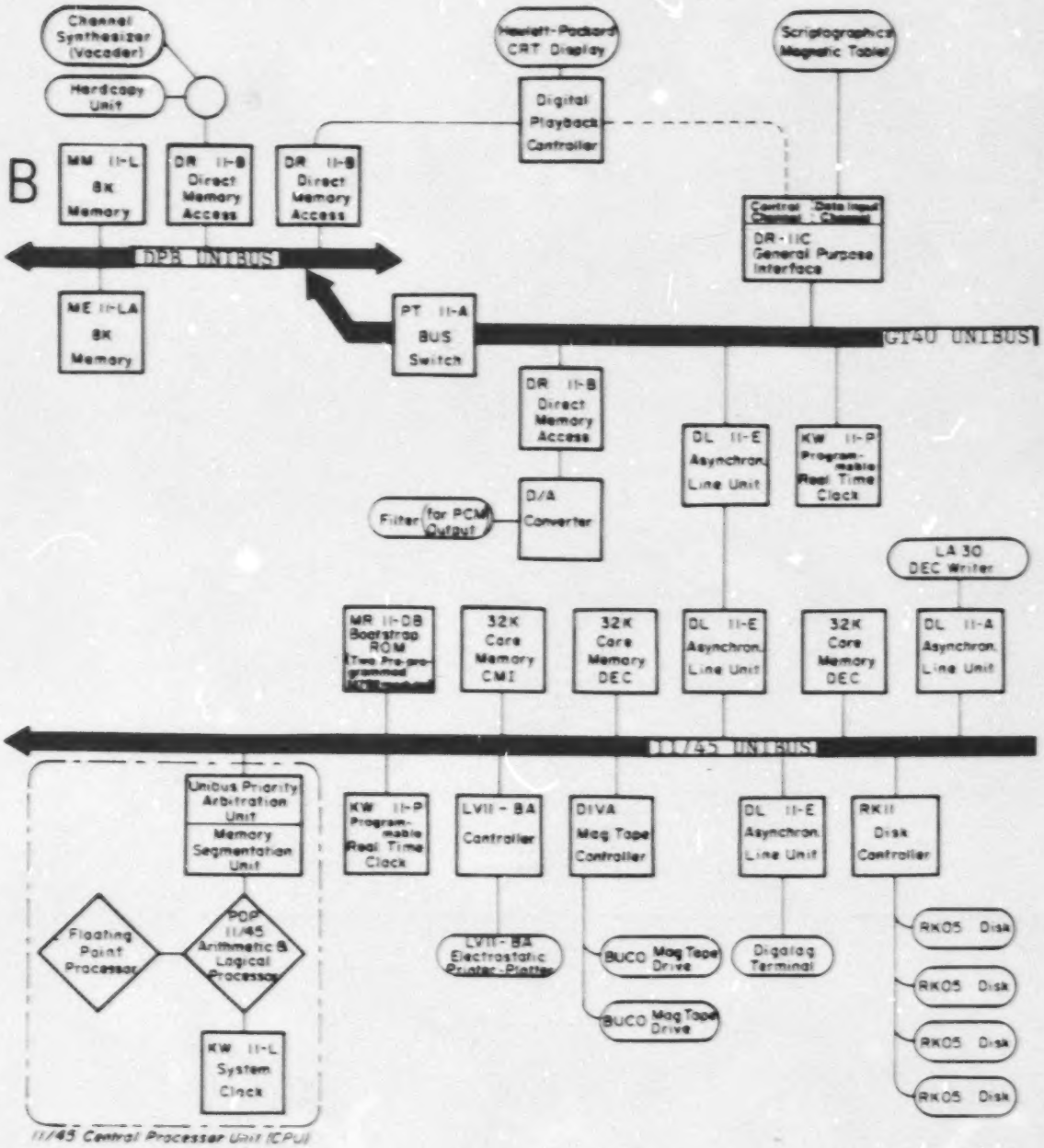


Figure 1: An overall view of the DPP showing the principal components available to the experimenter. The physical components actually handled by the user have their profiles depicted with heavy lines. Memory A stores the PCM data input via the A/D converter. Memory B is occupied by Spectrum and related graphic data, which are viewed via the HP display and its controller.


Channel Synthesizer (Vocader)
Hardcopy Unit
Hewlett-Packard CRT Display
Digital Playback Controller
Scriptographics Magnetic Tablet
B
MM 11-L 8K Memory
DR 11-B Direct Memory Access
DR 11-B Direct Memory Access
DR-11C General Purpose Interface
DPB UNIBUS
ME 11-LA 8K Memory
PT 11-A BUS Switch
GT40 UNIBUS
DR 11-B Direct Memory Access
DL 11-E Asynchron. Line Unit
KW 11-P Program-mable Real Time Clock
Filter for PCM Output
D/A Converter
LA 30 DEC Writer
MR 11-DB Bootstrap ROM
32K Core Memory CMI
32K Core Memory DEC
DL 11-E Asynchron. Line Unit
32K Core Memory DEC
DL 11-A Asynchron. Line Unit
PDP 11/45 UNIBUS
Unibus Priority Arbitration Unit
Memory Segmentation Unit
Floating Point Processor
PDP 11/45 Arithmetic & Logical Processor
KW 11-L System Clock
KW 11-P Program-mable Real Time Clock
LV11 - BA Controller
LV11 - BA Electrostatic Printer - Plotter
DIVA Mag Tape Controller
BUCO Mag Tape Drive
BUCO Mag Tape Drive
DL 11-E Asynchron. Line Unit
Digalog Terminal
RK11 Disk Controller
RK05 Disk
RK05 Disk
RK05 Disk
RK05 Disk
11/45 Central Processor Unit (CPU)

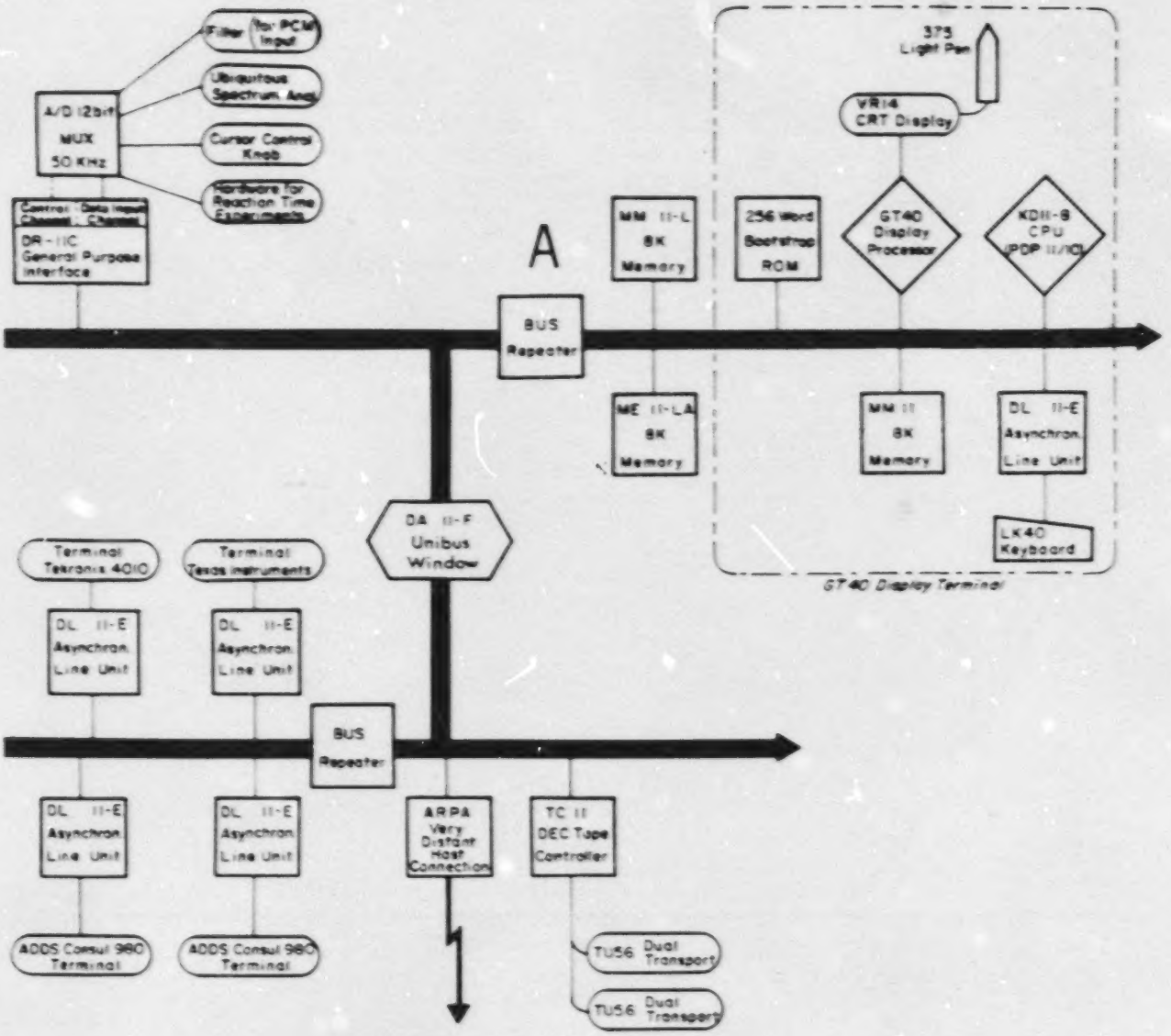


Figure 2: The GT40 display terminal (upper right-hand side) is connected to a unibus extension that supports the DPP display system. A PDP-11/45 (bottom left-hand side) is connected to the GT40 via a unibus and an asynchronous line interface.

and Analog-to-Digital (A/D) and Digital-to-Analog (D/A) conversion facilities. In addition, a bus switch provides access to a further unibus segment (DPB) and 16K of core memory (memory B), which is mapped into the same set of address locations as the auxiliary memory A connected directly to the GT40. When the bus switch is closed, the auxiliary memory is disabled and the CPU is allowed to access the memory connected to the unibus segment. Conversely, when the bus switch is open, the CPU may access the auxiliary memory block. The main purpose of the bus switch is not only to increase the memory capacity of the system by providing access to two memory units, but also, when the bus switch is open, to allow Direct Memory Access (DMA) devices connected to the unibus segment to operate independently of the GT40 CPU. A total of two such DMA units is connected to the unibus segment. One DMA device outputs digital data to either a hard copy unit or to a 40-channel synthesizer, while the other (the scan DMA) delivers data to a fast cathode-ray tube (manufactured by Hewlett Packard and termed the HP display), which is driven by a specially built Digital Playback Controller (the DPP controller). Commands to this controller are issued through an interface connected to the GT40 unibus.

The PDP-11/45 utilizes a multiuser operating system (RSX-11D) supplied by the computer manufacturer. Several keyboard terminals are interfaced directly to this computer, while additional terminal access is provided from the GT40 through a pair of asynchronous line interfaces connected back-to-back. Thus the keyboard associated with the GT40 has the capability of communicating directly with the PDP-11/45 operating system and of requesting that programs be loaded and executed. Programs or data to be loaded into the memory blocks of the GT40 pass through the unibus window. The latter is a two-way device that maps bus addresses in the region 96 to 112K on the PDP-11/45 unibus to memory locations 0 to 16K on the GT40 unibus, and, in the other direction, maps the region 24 to 28K on the GT40 unibus to any 4K area of the PDP-11/45 memory under program control. The bus window has the status of a peripheral device, and control over window functions is exercised through an RSX-11D compatible device handler. Using the interrupt mechanism of the unibus window, processes resident in the GT40 can be activated by programs running in the PDP-11/45.

STAGE 2: FUNCTION OF THE DPP HARDWARE

All special-purpose hardware devices associated with the DPP are either attached to the GT40 unibus extension or to the unibus segment DPB. The most important of these devices is the DPP controller and modified DMA device that drives the HP display used to exhibit spectrograms and some graphical data. One source of spectrographic data is a real-time spectrum analyzer that accepts speech input from a microphone, tape recorder, or synthesizer. The analyzer, having an 80-Hz frequency resolution, stores the spectrum data in either one or both of the two 8K blocks of core memory B attached to the unibus segment DPB. Concurrently, the same speech input signal is also filtered, sampled, digitized and subsequently stored as Pulse-Code-Modulated (PCM) data in the two auxiliary 8K blocks of core memory A attached directly to the GT40. The sampling rate for both the spectrum analyzer and the A/D converter is set at 10 kHz and is derived from a common clock. Prior to sampling, the speech is low-pass filtered with a cutoff frequency of 4.8 kHz and given high-frequency preemphasis. The latter is subsequently corrected by the output filter associated with the D/A converter.

The spectrographic data are stored in seven-bit integer form in successive bytes of the display memory. The eighth and most significant bit of each byte is reserved for an embossing function, to be described later. On output, the spectrographic data points are transmitted via the DPP controller to the HP display scope, creating a rectangular raster image of 128 x 128 elements scanned vertically--each successive vertical scan line representing a 12.8-msec analysis interval. Prior to generating a display, the controller is loaded with instructions specifying the starting point of the scan lines on the screen. The GT40 then loads the scan DMA with the starting address in memory for the spectral data as well as the range of data to be displayed. When the display is initialized, the scan DMA output is transferred to a portion of the controller that quantizes the spectral energy values into 13 distinct beam intensity levels. Alternate 128-byte sequences providing first the even scan lines and then the odd lines are output by the scan DMA to produce an interlaced raster that minimizes flicker. At the end of each raster frame, the scan DMA generates an interrupt and the next frame is initialized by an interrupt routine. The image seen by the experimenter consists of up to two spectrograms, lying one above the other (see Figure 1), which have an aspect ratio closely resembling that used by conventional sound spectrographs.

A short vertical line, whose horizontal position is controlled by a potentiometer, allows one to scan along the lower spectrogram and to indicate points of interest. This cursor is drawn on the screen in "emboss mode," using a special feature of the DPP controller that outputs at maximum intensity all spectrum bytes that have their eighth bit set. The mode can be canceled by simply removing the eighth bit, which leaves the remaining seven-bit data field intact and available for display. To update the position of the cursor, the potentiometer output is sampled at each execution of the display interrupt routine and interpreted as a scan line number. If the cursor line number changes between successive display frames, the old cursor bytes are located, their eighth bits are erased, and new bits are set in the current line prior to reinitialization of the display. The cursor is used to locate sections of the spectrum and associated PCM data fields for analysis purposes.

The vocoder synthesizer attached to the DPB unibus segment is a 40-channel ringer type with alternate channels coupled in opposing phase. Output from the spectrum display memory B via a DMA unit consists of a sequence of 128 data bytes per scan. Forty of these samples are selected by the vocoder hardware at roughly equal spectral intervals in the 5-kHz range and are held in data registers. These registers modulate the output amplitude of the 40 resonators, each excited by a noise or buzz source. Both the frequency and amplitude of the latter source can be controlled by data inserted manually with the writing tablet or, alternatively, obtained by computation. These values are stored with their associated spectrogram in the spectrum display memory.

STAGE 3: THE DPP SOFTWARE

The basic aim of the DPP software is to provide the investigator with a range of programs enabling him to display and manipulate speech data. This is achieved through the use of a nuclear group of programs operating in the GT40 and a collection of more complex data processing algorithms that are executed in the PDP-11/45. The unifying structure coordinating access to all of these programs is provided by a program called the "Supervisor." Commands issued to

the Supervisor can initiate programs in either computer, and additional programs designed to manipulate speech data can be written for whatever computer is best suited to the particular task.

The relationships among the Supervisor, the operating system of the PDP-11/45, the communication pathways to the GT40, and the main components of the DPP are shown in Figure 3. The software supporting the GT40 keyboard permits it to be switched into either of two modes. In local mode, requests for Special Functions (SFs) issued by a user are interpreted by a local command parser in the GT40 and executed directly by that processor, while in remote mode, user-commands are routed to the PDP-11/45 operating system through which the Supervisor is accessed. User-Defined Programs (UDPs) installed in the PDP-11/45 perform processing operations that require significant amounts of computation. They are usually written in FORTRAN but may be prepared in other languages. Examples are calculations of the Fast Fourier Transform (FFT), Linear Predictive Coefficients (LPC), and fundamental frequency functions from PCM speech data. The Supervisor provides user access to these programs through a command interpreter that recognizes a three-character name typed by the user and initializes the appropriate routine. Access, on the part of UDPs, to data stored in the GT40 memory units is obtained through the unibus window and its software handlers. In addition, UDPs installed in the PDP-11/45 may use any of the SFs resident in the GT40 merely by issuing a command and initiating an interrupt through the unibus window. In most cases, when the GT40 SF has completed execution, a signal is transmitted back to the PDP-11/45 and the initiating UDP then resumes. However, some UDP and SF pairs are designed to run concurrently. This feature allows long periods of speech to be sampled (see description of the TALK UDP below).

Speech spectrum samples and other data retained by mass storage devices attached to the PDP-11/45, can be transferred to the 16K block of display memory B connected to the GT40's unibus segment. The remaining auxiliary memory block A usually receives the associated PCM data. However, memory A can serve as a data sanctuary into which spectrographic and other data can be temporarily stored.

SPECIAL FUNCTIONS

When operated in remote mode, the GT40 keyboard provides access to the entire range of UDPs and SFs by communication with the Supervisor. In local mode, the keyboard gives the user control of a basic set of SFs that can be accessed directly. These programs are initiated by single letter commands issued at the GT40 keyboard. A full list of these commands is made available to the user in table form on the GT40 display in response to the letter D.

Examples of some of the SFs installed in the GT40 are programs that:

1. generate the spectrogram and PCM samples for a 1.6-sec passage of speech,
2. plot the speech waveform on the GT40 display from data in memory A,
3. plot a spectral cross section on the HP display corresponding to a cursor-selected point on the spectrogram display,
4. print a hard copy of the HP display,

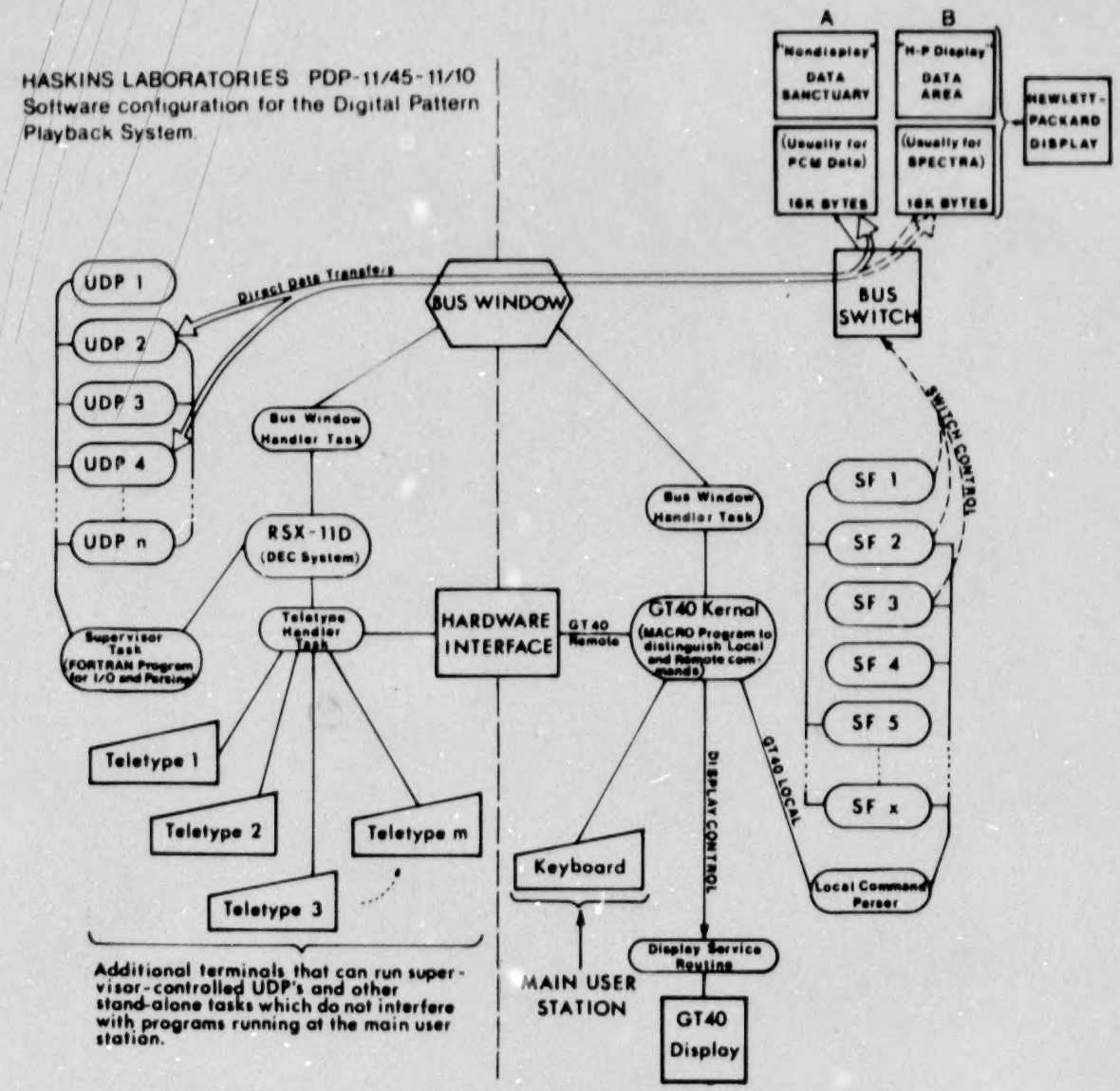


Figure 3: The software of the DPP is installed partly in the PDP-11/45 (left-hand side of diagram) and partly in the GT40 (right-hand side). Executive control is exercised by the supervisor task, which is installed in the PDP-11/45. The experimenter communicates with the supervisor from the main user station via the hardware interface. Data for display by the DPP are transferred directly via the bus window from mass storage devices connected to the PDP-11/45.

5. input data from the writing tablet,
6. transfer speech data between A and B memory blocks,
7. output speech data to the vocoder or D/A converter,
8. remove or insert interpolated spectral scan lines.

USER DEFINED PROCESSES

A singular feature of the DPP is its extensible software structure. Key components of this structure are the UDPs, which are executed by the PDP-11/45 under the control of the Supervisor program. At the present time about 10 UDPs have been written and installed.

The DUMP UDP is a program that moves 8K blocks of data, usually PCM and spectrum data, between memories A and B and file-structured mass storage devices connected directly to the PDP-11/45 unibus. As parameters, DUMP requires the file name, an indication of whether the file is to be read or written, and four digits indicating the order in which the four 8K blocks of the GT40's A and B memories are to be accessed. When one reads a file, he must specify the starting record for each block to be filled, or a set of default values will be used to fill all four 8K blocks from the file in sequence. DUMP can be used to save the contents of the A and B memories at the end of a work session and to restore them later. The program is also useful for storing the synchronized PCM and spectrum data that have been captured and stored in the A and B memories by a special function in the GT40.

To retrieve and display spectrum tokens that occupy no more than 16K bytes together with their corresponding PCM files, the DISPLAY UDP is used. This program also outputs to the HP display, labeled plots of a token's amplitude, fundamental frequency, or other functions that have been computed by other UDPs and stored on a disk. A file name is the only essential parameter required by the DISPLAY UDP. However, if different default values are required, the starting and ending file record numbers, the display frame, or memory block in which the display should appear, and the y-scan , or starting point within the display frame, may be provided. Also, plot scale parameters may be specified for function files. Since the DISPLAY UDP remembers the parameters that have been typed in each time it is called and uses these previously given parameters as default specifications, it is possible to retrieve and display corresponding segments of synchronized PCM, spectrum, fundamental frequency, and amplitude files by typing a few short commands.

One of the most frequently used UDPs is called PITCH. By an adaptive autocorrelation method, this program computes a fundamental frequency function for PCM data stored in a disk file or in the GT40 memory A (see Figure 3). The resulting frequency function is stored on disk and optionally plotted by the DISPLAY UDP. The file containing the frequency function may be saved for later reference or to control the fundamental frequency of the channel synthesizer (vocoder) and to provide voiced-voiceless switching.

When corresponding spectrum and fundamental frequency files are available on disk, the VOCODER UDP can be used to calculate control files containing the buzz, hiss, and fundamental frequency parameters for the synthesizer. The data in these files can later be transferred to the proper memory addresses using the DISPLAY UDP, DUMP UDP, or VDUMP--a special UDP for conveying switching

information to the appropriate memory addresses. When a spectrum is already installed in memory and the corresponding fundamental frequency file is on disk, the VOCODER UDP can calculate the required switching information and deposit it in memory in one step. The current algorithms for determining buzz and hiss rely on the fundamental frequency data to locate voiced and unvoiced portions of speech. The buzz parameter is set for all scans where voice pitch is detected. Occurrences of unvoiced hiss are located by summing the spectral values above 1000-kHz in each y-scan where voice pitch cannot be detected and comparing the result with a threshold value that may be input as a parameter. The minimum and maximum fundamental frequency values of the synthesizer may also be input as parameters. A test for voiced-hiss is being developed.

To date, UDPs that calculate four types of loudness-function from spectral data are available. The AMPLITUDE UDP is the least embellished of these functions. AMPLITUDE requires a spectrum file specification or percent symbol (indicating that the input data is located in the B memory). The function is computed by summing the squares of the spectral amplitude values contained in each y-scan, then normalizing and converting the result to decibels. The result is stored in a disk file and optionally plotted by the DISPLAY UDP. The LOUDNESS UDP accepts two additional parameters: a code for a frequency weight function and a code for smoothing along the time axis. Again, the result is stored in a disk file ready for plotting.

PCM samples comprising tokens of 1.6-sec duration are the largest that can be captured in the GT40 A memory blocks. The TALK UDP is designed to input longer passages of continuous speech. TALK is actually a pair of routines, a UDP in the PDP-11/45 and an SF in the GT40, for inputting, outputting, and editing (that is, separating into RSX-11D format files) long streams of PCM data sampled at a 10-kHz rate. By double-buffering and using a scratch disk as an interim PCM storage device, slightly over 2 min of data sampled at 10 kHz can be input. In the output mode, the user specifies the disk blocks to be played and can start and stop output by raising sense switches. The file editing mode permits the deletion of unwanted data and provides an opportunity to move data from the scratch disk to a file structured volume. This latter step is necessary because the FORTRAN file service routines are too slow to keep in step with a 10-kHz PCM sampling rate. Once the data are available in RSX-11D file format, they can comprise the input for other UDPs including the FFT UDP, which can be used to calculate a spectrogram directly and can provide a more accurate, although slower, alternative to the hardware analyzer.

APPLICATIONS OF THE DPP

Many basic studies of the cues that help the listener distinguish phonetic segments require the manipulation of a particular acoustic feature and the assessment of its effects by listener tests. The DPP provides a variety of operators useful in research of this kind and some examples will serve to illustrate the point.

One potential group of experiments arises from the ability to input real speech, modify certain features, and resynthesize the signal via the vocoder. Thus, the prosodic features (fundamental frequencies, durations, and amplitudes) of natural speech passages can be selectively altered to explore their role in the process of speech perception. For example, questions concerning the effect

of speaking rate on the perception of voice-onset time in initial stop consonants can be explored. Here the technique might start with a naturally spoken sentence containing a stop consonant and use the facilities to alter the overall speech rate without increasing vocal pitch, to vary the durations of individual vowel segments or transitions preceding the consonant, or to modify the pitch or amplitude contours. An important common factor in all of these experiments is that they are performed with stimuli whose origin is natural speech. This means that they will contain naturally coarticulated segmental features free from many potential artifacts of synthetic speech while providing precise control over the parameters of interest and retaining the flexibility and reproducibility of the synthetic medium.

A second group of applications is concerned with the development of methods for automatic speech analysis in speech understanding systems. At the primary levels of analysis, the incoming acoustic signal must be segmented and the constituent features of each segment must be identified. The choice of these features is of critical importance in selecting the best strategy for using a feature labeled dictionary. Not all of the features that can be identified are necessarily useful. Some are simply redundant and supply no additional information, and, among the remaining features, not all are equally reliable because of wide variations in prominence from speaker to speaker. In these circumstances, feature evaluation studies are necessary to eliminate the redundant features and to establish reliability estimates for the rest. These estimates may then be employed in dictionary search strategies in which the spoken words are sought in terms of feature vectors.

One method of feature evaluation to which the DPP is well-suited involves human subjects in the task of identifying selected sets of features and using them in a dictionary search procedure. A sentence is selected and spoken by the experimenter and input to the DPP. The subjects, who do not hear the sentence, may retrieve and display it in spectrographic form. Their task is to divide the sentence into segments, classify the features according to their individually preferred schemes, and then seek matching spectrograms from a dictionary of reference items containing the required words among many similar items. The contents of the dictionary are classified according to the subjects' personal schemes in preliminary sessions. Each classification scheme may then be evaluated in terms of the number of potential matches returned on each dictionary enquiry, the number of such queries, and the accuracy of the subjects' matching performances.

CONCLUSIONS

This report describes the principal facilities currently offered by a combination hardware and software device called the digital pattern playback. The inherent extensibility of the DPP software design means that its present complement of facilities is not intended to remain unchanged and that this paper is necessarily an interim description. For the future it is planned that programs to perform synthesis by rule will be made available to use either the OVE III synthesizer or a software synthesizer that can be configured as a serial or parallel formant resonance system. It will then be possible to compare readily the spectra of synthetic and natural speech as well as to create stimuli containing segments from both sources. In addition, a computer model of the vocal tract, planned as a part of the Laboratories' future

research, may use the display facilities of the DPP—particularly the GT40 display—to display cross sections of the vocal tract. User interaction by means of the light-pen offers an opportunity for the experimenter to manipulate the shape of the model vocal tract (as he can now manipulate the shape of the spectrum), calculate the resultant acoustic signal, and listen via the D/A output system.

The present form of the DPP is an outgrowth of the work of the Haskins Laboratories staff and has received the benefit of suggestions and contributions from many of its members. Its easily extensible program structure facilitates the addition of new SFs and UDPs. Without requiring a detailed knowledge of the entire system, an investigator can write routines in a higher-level language (e.g., FORTRAN) and thereby extend the collection of processing tools available to future users. It is therefore expected that the DPP will become very much a shared facility not only in use but also in design.

REFERENCES

Cooper, F. S. (1953) Some instrumental aids to research on speech. In Proceedings of the 4th Annual Round Table Meeting on Linguistics and Language Teaching. (Washington, D.C.: Georgetown University), pp. 46-53.

In (Qualified) Defense of VOT*

Leigh Lisker+

ABSTRACT

The VOT measure has been said to provide the single most nearly adequate physical basis for separating homorganic stop categories across a variety of languages, granted that other features may also be involved. That transition duration affects perceived voicing of synthesized initial stops of one specific language, English, has suggested the hypothesis by Stevens and Klatt that a detector responsive to rapid formant frequency shifts after voice onset better explains the child's acquisition of the contrast than does some mechanism which responds to VOT directly. If such a detector is part of our biological equipment, then it seems remarkably underutilized in language, for the hypothesis asserts that basic to voicing perception is whether laryngeal signal is or is not present during the interval in which the stop-vowel shift occurs. In effect, the "archetypical" voiceless stop is aspirated. Not only do many languages not possess voiceless aspirates, but even in English aspiration is severely restricted. Of course the VOT measure has its limitations--it is inapplicable to prepausal stops. However, there are much more serious difficulties with the posited detector, since even for the English initial stops there is evidence that the presence of a voiced first-formant transition is not required for /b,d,g/, nor is its absence necessary for /p,t,k/, provided appropriate values of VOT are set.

To judge from much current discussion of the cues to stop voicing, it appears that voice onset time (VOT) is taken quite simply as the duration of the time interval between onset of the release explosion, or burst, and the onset of glottal pulsing, two points readily discoverable in the acoustic representation of stop-vowel sequences. It is one of several possible measures by which to express the temporal relation between the laryngeal and supraglottal maneuvers involved in producing these sequences. In initial position the choice of voicing onset as the acoustic index to the timing of the laryngeal action is the only practical one; burst onset as the acoustic marker of the supraglottal stop articulation is not the only possible one, but it is the one most easily located by eye in either the spectrographic or oscillographic display of the speech

---

*To appear in the Proceedings of the 8th International Congress of Phonetic Sciences.

+Also University of Pennsylvania, Philadelphia.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

signal. Measurements of extensive samples of speech have suggested that the VOT measure provides the single most nearly adequate physical basis for separating the homorganic stop categories of English and a number of other languages (Lisker and Abramson, 1964). At the same time they have shown that VOT values for a given stop category are not completely independent of certain other factors. For English, for which we have greatest quantity of data, knowing the VOT value for an initial stop-vowel combination is not quite enough to assign the stop to one or the other of the two sets /b,d,g/ and /p,t,k/; for some VOT values it is necessary to have information as to place of stop closure, degree of stress on the syllable, and perhaps some other factors as well (Lisker and Abramson, 1967). If we can justifiably assume that these stops are phonetically unambiguous, then clearly the listener attends to some feature or features in addition to the absolute duration from burst to voicing onset. Of course it is more often the case that the absolute VOT value will suffice for the correct classification of a stop, so that it may be that for such stops VOT is the paramount, and possibly the sole, cue for the listener. We cannot decide whether this is true by observing natural speech, since the several acoustic consequences of varying the timing of vocal fold adduction and onset of oscillation (that is, the "underlying" VOT dimension) do not vary in a mutually independent manner in natural speech. We must turn to speech synthesis, gaining the advantage of a more tractable speech source at the cost of an acceptable loss in naturalness.

Experiments in the perception of synthetic speech patterns have yielded results consistent with findings for natural speech, and they show that listeners may not base their labeling judgments of initial stops exclusively on the duration of the interval between the burst and voicing onsets. Thus, if the first formant is not considerably attenuated in this interval, then, even if that interval has a duration appropriate to /p,t,k/, listeners will be uncertain in their responses. Moreover, variation in the fundamental frequency contour following voice onset may also affect those responses (Haggard, Ambler, and Callow, 1970; Fujimura, 1971). Particularly because of the first-formant attenuation, the so-called F_1-cutback, some of us at Haskins Laboratories have from the beginning of our studies of VOT thought of this as an articulatory dimension; we have even spoken of a VOT continuum, a term quite inapplicable, in any acoustic sense, to the synthetic speech stimuli used in our experiments in the perception of stop voicing. Thus F_1-cutback, F_0 contour variation, and any other acoustic consequences of a variation in the timing of laryngeal action, are all, in this view, components of the VOT dimension. Two of these--the timing of voice onset and attenuation in the frequency region of the first formant during the interval between burst and voicing onset--are of major perceptual significance. One feature that has been under study recently--the duration of the first-formant transition--is of particular interest because it significantly affects judgments of stop voicing (that is, identification as /b,d,g/ or /p,t,k/) and yet appears to be independent of the larynx. Stevens and Klatt (1974) have shown that the VOT boundary between synthetic English /da/ and /ta/ varies with the transition duration. This effect is exemplified by the data displayed in Figure 1, which replicate the finding reported by Stevens and Klatt. In this experiment, 20 subjects provided four responses each to each of 56 stimuli involving VOT values from +5 and +65 msec and transition durations ranging from +20 to +115 msec. The 50 percent crossovers between /da/ and /ta/ responses shift from a VOT value of about +21 msec, for the shortest transition, to one of about +48 msec, for the longest. On the strength of this sort of finding Stevens and Klatt have suggested that listeners may categorize a stop on the

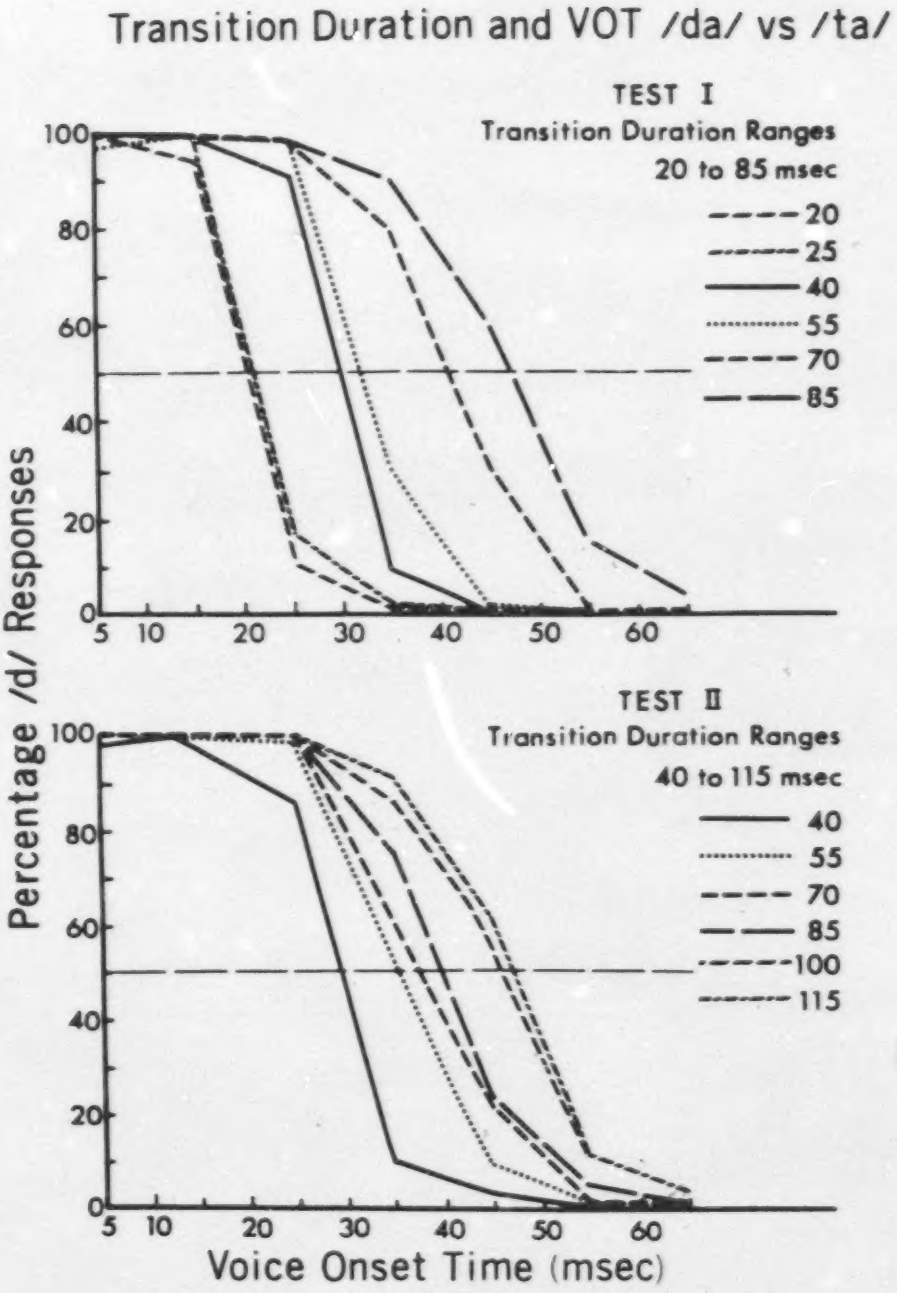


Figure 1: Response data in a forced-choice labeling test of /da/-/ta/ patterns synthesized on the Haskins Laboratories parallel formant synthesizer. Each data point represents 4 judgments by each of 20 Ss.

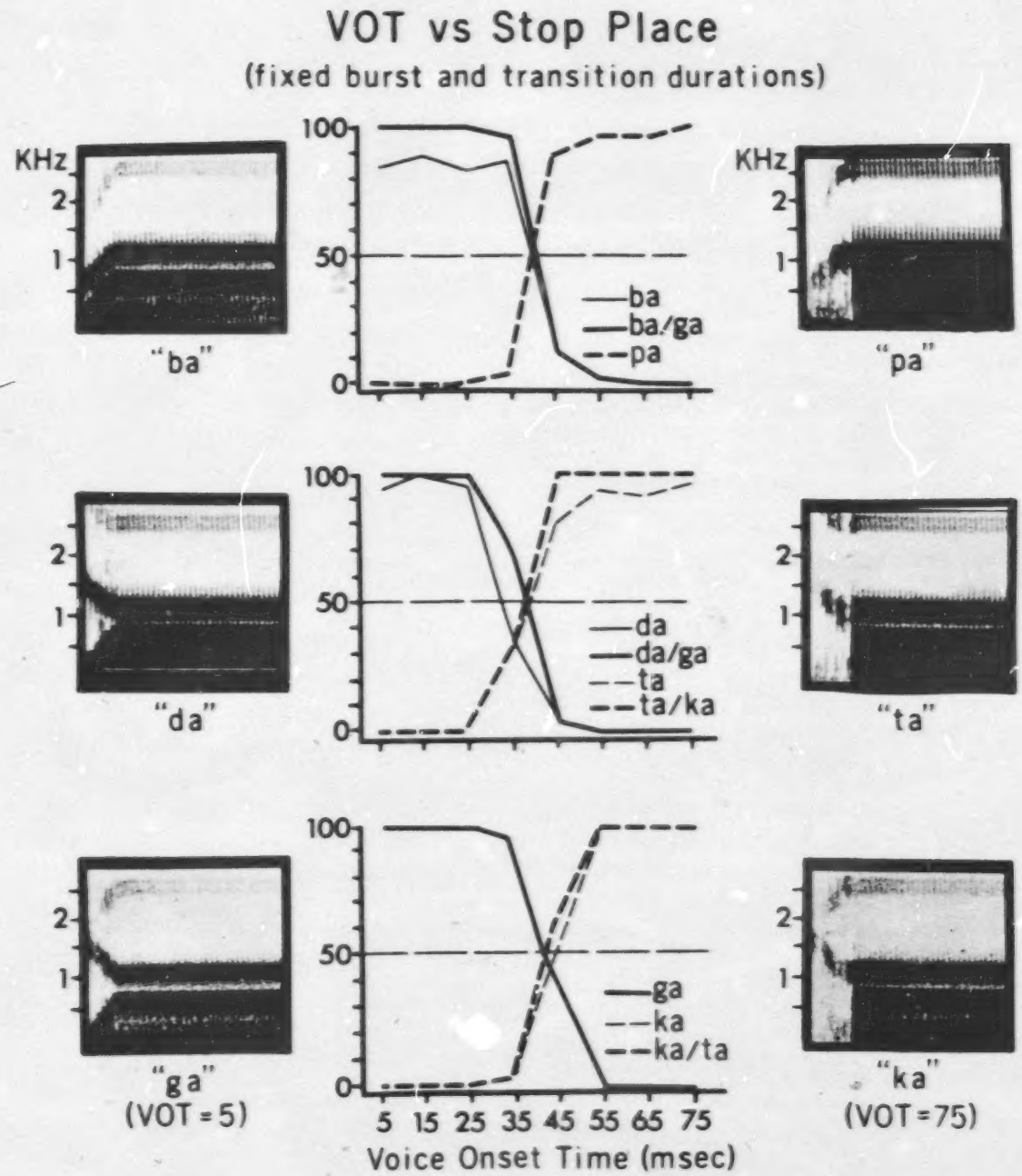


Figure 2: Labeling response data from 8 Ss (4 trials) required to classify perceived initial stop in each of 24 stimuli (8 VOT values × 3 transition configurations) with respect to voicing state and place of articulation.

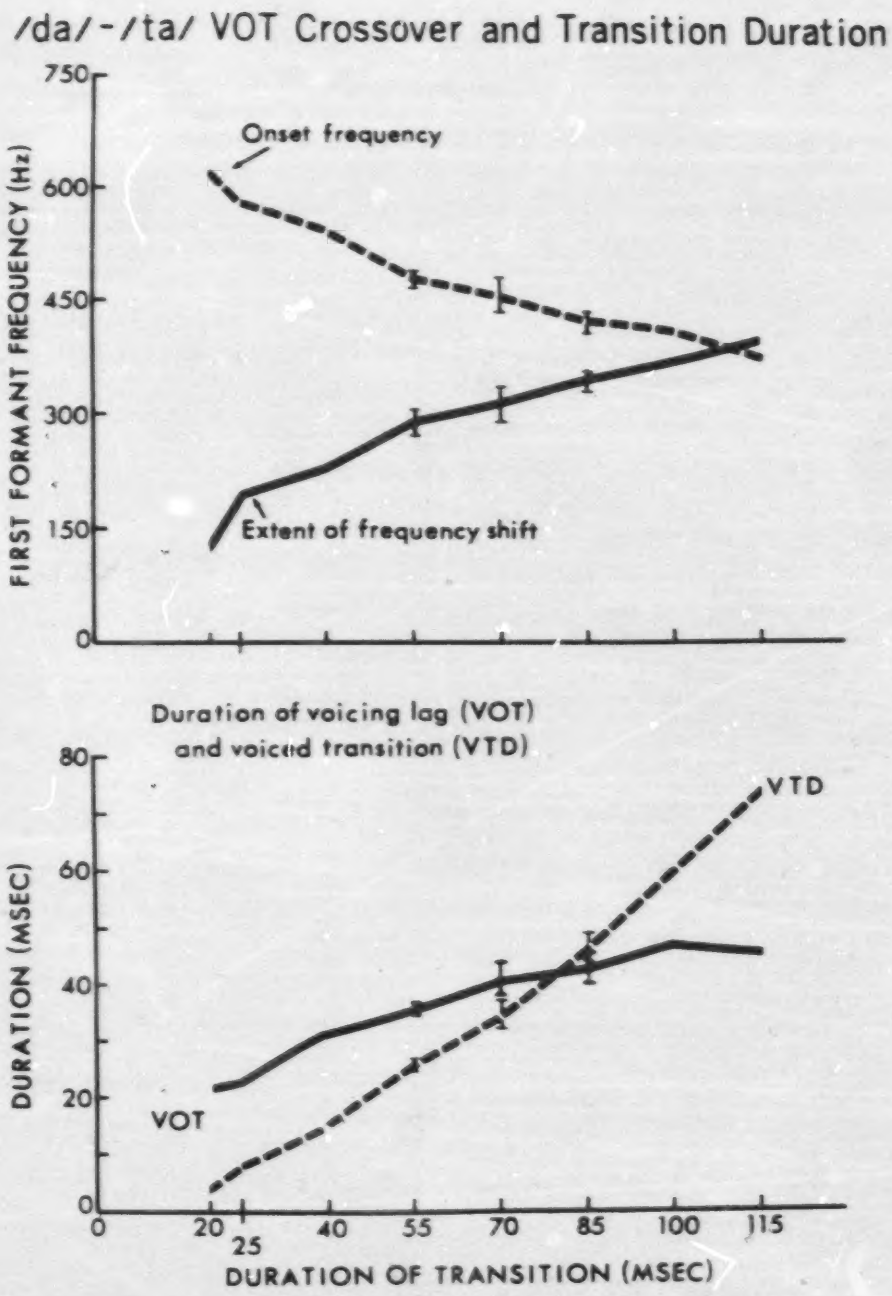


Figure 3: The 4 curves are derived from the same data that are represented in Figure 1. The spectrograms represent minimum and maximum VOT values tested.

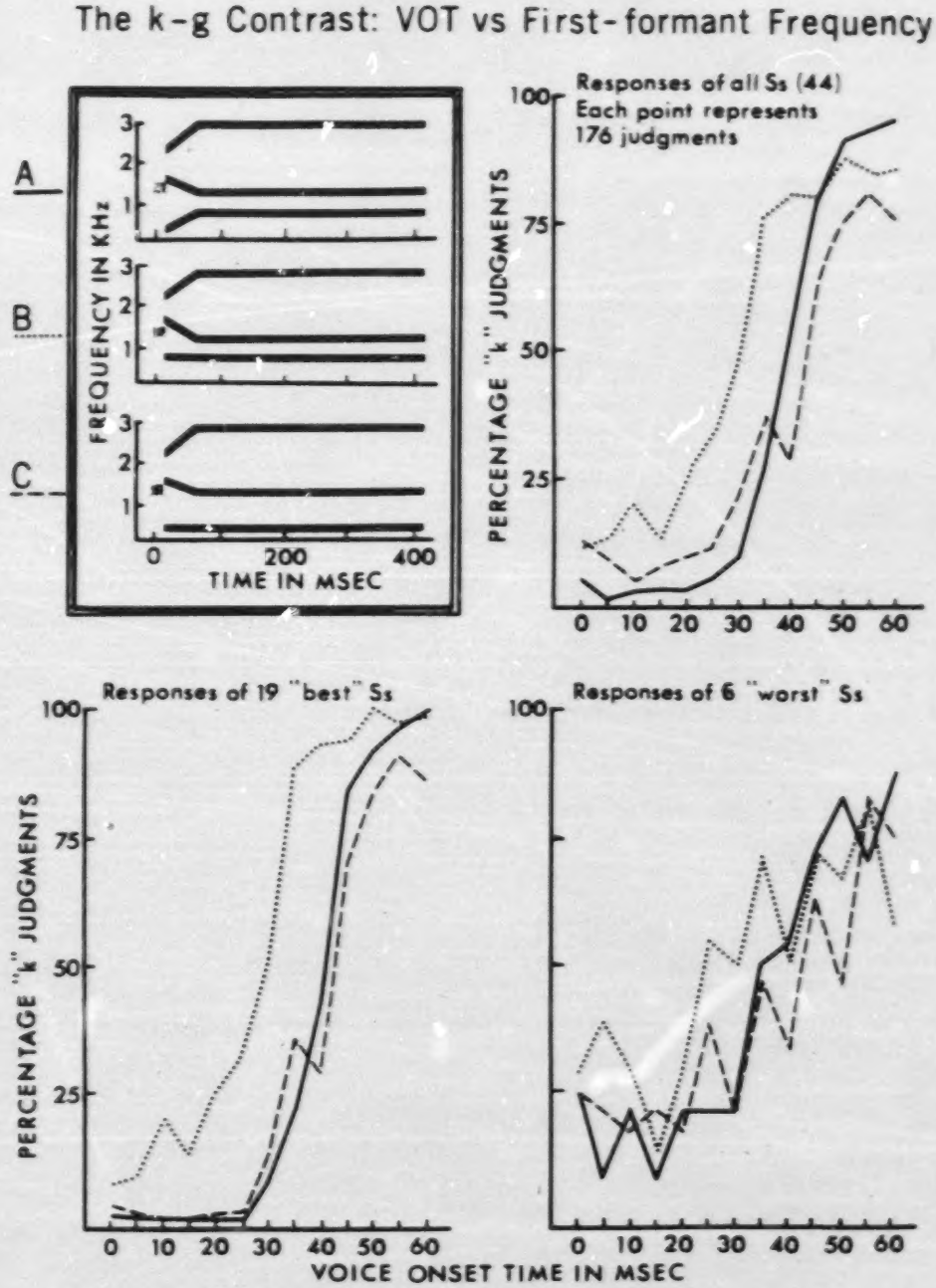


Figure 4: Responses of 44 Ss to patterns of the types shown. "Best" Ss are those showing the greatest consistency in behavior; the "worst" Ss responded identically on four trials to the smallest percentage of the 39 acoustically distinct stimuli presented.

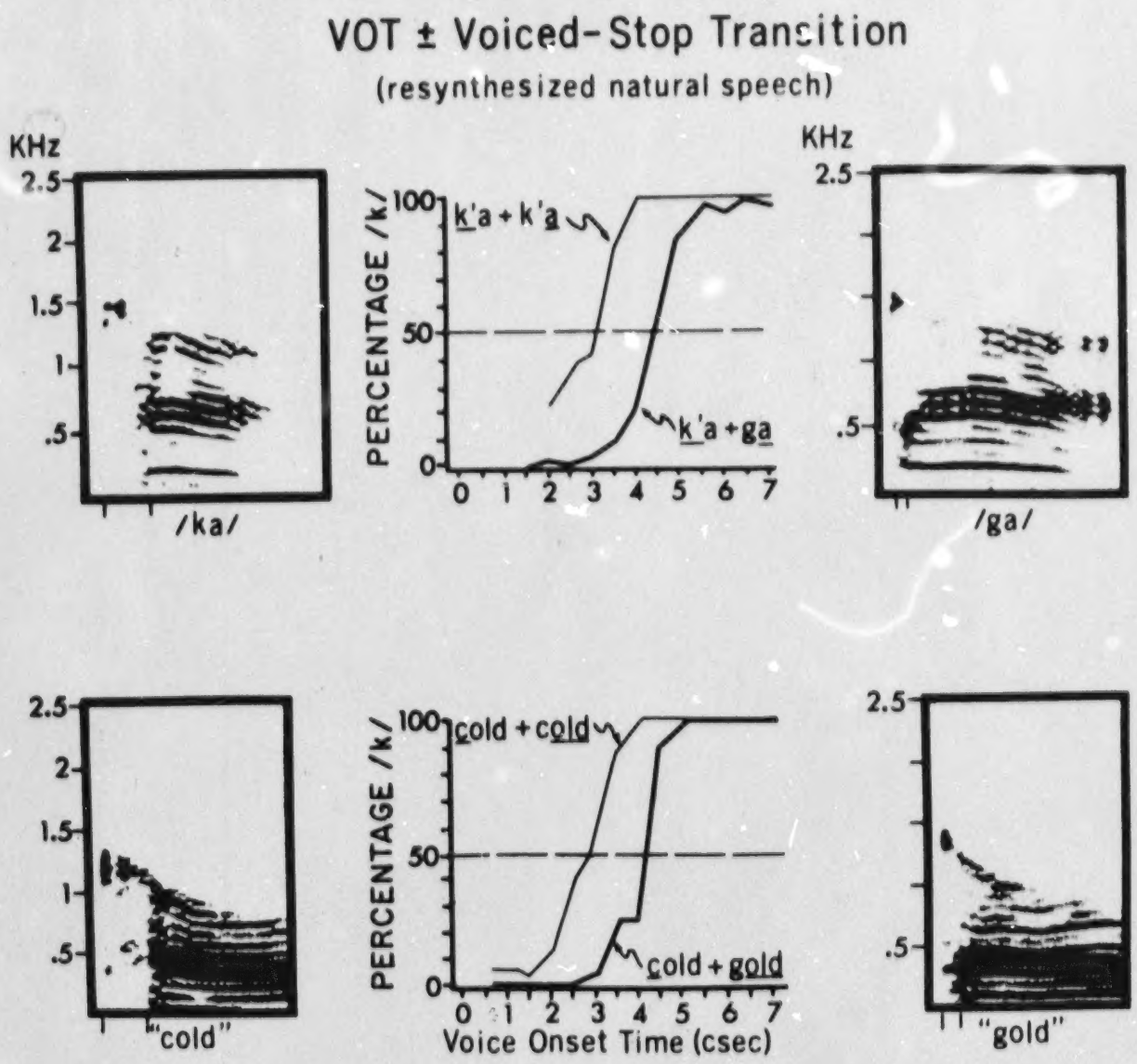


FIGURE 5

Figure 5: The spectrograms are of naturally produced syllables, the nonsense pair /ka/ and /ga/ and the minimal word pair cold and gold. Test stimuli were prepared by combining the /k/ burst and varying durations of aspiration with the voiced portions of both members of each syllable pair. Editing was done electronically with the Haskins Laboratories pulse-code-modulation (PCM) system.

basis of the timing of voice onset relative to the end of the formant transition to the following vowel rather than to the burst onset. They point out that this suggestion would explain the place effect on the VOT boundary; for its progressively greater value as one goes from labial to alveolar to velar place is said to be matched by increasing durations of the interval between burst onset and achievement of the following steady-state vowel; the durations of the voiced first-formant transition ought to be more nearly the same for the three places of stop articulation. In fact, if stops of the three places are synthesized using the same durations of burst and first-formant transition, the VOT crossovers do not show a regular increase in the crossover values as the place of closure moves from front to back in the mouth. The data given in Figure 2 show no regular shift in VOT crossover; values for labial and velar places are the same, and the /da/-/ta/ crossover is slightly earlier. In one of the earliest Haskins experiments (Liberman, Delattre, and Cooper, 1958), the finding was quite the reverse, with the apicals showing the greatest value at the boundary; so that if these are significant differences, their basis is at present unclear. In any case, these data seem to support Stevens and Klatt (1974) in connecting the place effect observed in natural speech with differences in burst and transition duration that have been reported. But it is not at all certain that the data showing an effect of transition duration on the VOT boundary support their hypothesis that there is a first-formant transition detector whose response to a frequency shift following voice onset is a more reliable basis for a voiced-stop judgment than VOT. The display in Figure 3 represents the same data presented in Figure 1; the lower panel provides a comparison of the VOT boundary as a function of transition duration with the duration of the voiced first-formant transition. It would appear from this that the latter measure is rather more susceptible to variation with changing transition duration than is the VOT measure. The upper panel shows two additional features that may, on further study, turn out to be significant to stop voicing perception.

The notion of an F_1 transition detector implies that the presence of voiced F_1 transition is a stronger cue to voicing than a short voicing lag, and that the absence of a voiced formant transition is of greater importance perceptually than a long lag in voicing onset. In Figure 4 we have data from synthesis indicating that the absence of first-formant movement is not incompatible with voiced stop judgments, provided the VOT value is right. Those data further suggest that the onset frequency of the first formant plays some role in the matter. In Figure 5, data are given showing the labeling responses to stimuli derived from natural speech samples. They indicate that the presence of first-formant transitions produced by voiced velar articulations is not incompatible with /k/ responses, provided again that the interval between burst onset and the onset of voicing is long enough. If the presence of a voiced first-formant transition is neither necessary nor sufficient to provoke voiced stop judgments, while the absence of the same feature is neither necessary nor sufficient to induce voiceless stop judgments, then it is hard to entertain seriously the hypothesis of an F_1 feature detector that operates more reliably than even the simplest conceivable device responding to the absolute duration of the interval between burst and voicing onset.

REFERENCES

Fujimura, O. (1971) Remarks on stop consonants: Synthesis experiments and acoustic cues. In Form and Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen, ed. by L. L. Hammerich, R. Jakobson, and E. Zwirner. (Copennagen: Akademisk Forlag), pp. 221-232.

Haggard, M., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. J. Acoust. Soc. Am. 47, 613-617.

Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. Lang. Speech 1, 153-167.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English stops. Lang. Speech 10, 1-28.

Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653-659.

The Coarticulation of Tones: An Acoustic Study of Thai*

Arthur S. Abramson+

ABSTRACT

Phonologically distinctive tones have glottal repetition rate as their primary articulatory base. In many tone languages, F_0 contours derived from productions of isolated monosyllables are considered the ideal forms of the tones. These ideal contours are not fully realized in running speech. This may be ascribed to rhythmic conditions and to laryngeal coarticulation as a function of the immediate tonal environments. Ideal F_0 contours for the five tones of Thai were obtained for four native speakers. Then, in a neutral frame, a mid tone at the beginning and a mid tone at the end, the four speakers produced all possible sequences of two tones in sentences. The data yield the following effects: (1) embedding a tone in running speech causes some departure from its ideal contour; (2) laryngeal coarticulation is governed by the specific tonal context; (3) tonal contrast is not lost.

Phones are subject to so much coarticulatory perturbation that phoneticians and psychologists (MacNeilage, 1970) have speculated about the nature of the invariance supposedly present in any set of speech sounds said to manifest an underlying phonological entity. The question seems to have received relatively little attention in the matter of phonemic tones that have glottal repetition rate as their primary articulatory base. It is conceivable that in something as continuously variable as the pitch of the voice, the lexical tones undergo so much sandhi and other perturbations that the tone system is not fully preserved

---

*To appear in the Proceedings of the 8th International Congress of Phonetic Sciences, Leeds, England, 17-23 August 1975.

+Also University of Connecticut, Storrs.

Acknowledgment: While most of the analysis of data was performed at Haskins Laboratories, the data themselves were collected while the author was on sabbatical leave in Thailand on research fellowships from the American Council of Learned Societies and the Ford Foundation Southeast Asia Fellowship Program. I gratefully acknowledge the hospitality of Dr. Udom Warotamasikkhadit, Dean, the Faculty of Humanities, Ramkhamhaeng University, and Mrs. Mayuri Sukwiwat, Director, the Central Institute of English Language, both in Bangkok.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

in running speech. Even if this is not so, are the ideal fundamental frequency (F_0) contours, epitomized by citation forms, preserved in a wide variety of contexts? Very little instrumental work on this topic has appeared in the literature. Examples are an investigation of Vietnamese by Han and Kim (1974) and a brief study of Thai by Palmer (1969).

The language chosen for this study is Central Thai (Siamese), which has five phonologically distinctive tones. In earlier work (Abramson, 1962), ideal F_0 contours for these tones in citation forms were derived instrumentally and synthesized for perceptual validation. Although recent work (Erickson, 1974) lends general support to the curves found, there is considerable disagreement about the preservation of these forms and the system of tonal contrasts in running speech (Henderson, 1949; Gandour, 1975).

PROCEDURE

All possible sequences of two tones from the five tones of Thai--mid, low, high, falling, and rising--were embedded on monosyllabic words in a carrier sentence beginning and ending on the mid tone to form 25 grammatical sentences. The mid tone provided as neutral a frame as possible.[1] The list of sentences was recorded at normal conversational speed on different occasions by each of four native speakers, three women and one man. Given the usual artificiality of such recording sessions, the productions were judged natural by the speakers themselves and the experimenter. In addition, careful listening revealed no loss of tonal contrast for the embedded key words. Citation forms of the tones on isolated words were also obtained from the four speakers.

Patterns of F_0 were extracted from all recordings by means of Lukatela's (1973) computer-implemented autocorrelation method. The resulting graphs permitted the examination of every tone in each left and right environment along the time axis.

RESULTS

Although all the data will not be presented here, I hope to make a full analysis available in subsequent reports. Selected environments for speaker P.C., one of the women, will be shown. They are rather representative of all four speakers.

In Figure 1 we see P.C.'s citation forms. The F_0 contours are normalized in time and displayed as percentages of her total voice range. The mid, low, and high tones are often said to be static or level tones, while the falling and rising tones, which have larger movements through the voice range, are called dynamic or contour tones. The other speakers have similar citation forms except that for two of them the mid tone slopes downward a little more and, for one of them, the falling tone starts its sharp drop immediately.

Turning to an examination of the tones and their contexts in the sentences, we find average curves for all of P.C.'s tones in the environment of the following mid tone in Figure 2. The left-hand portion of the figure shows the full set of tones in its environment, while the right-hand portion shows, as indicated

---

[1] The sentences were taken from Palmer (1969).

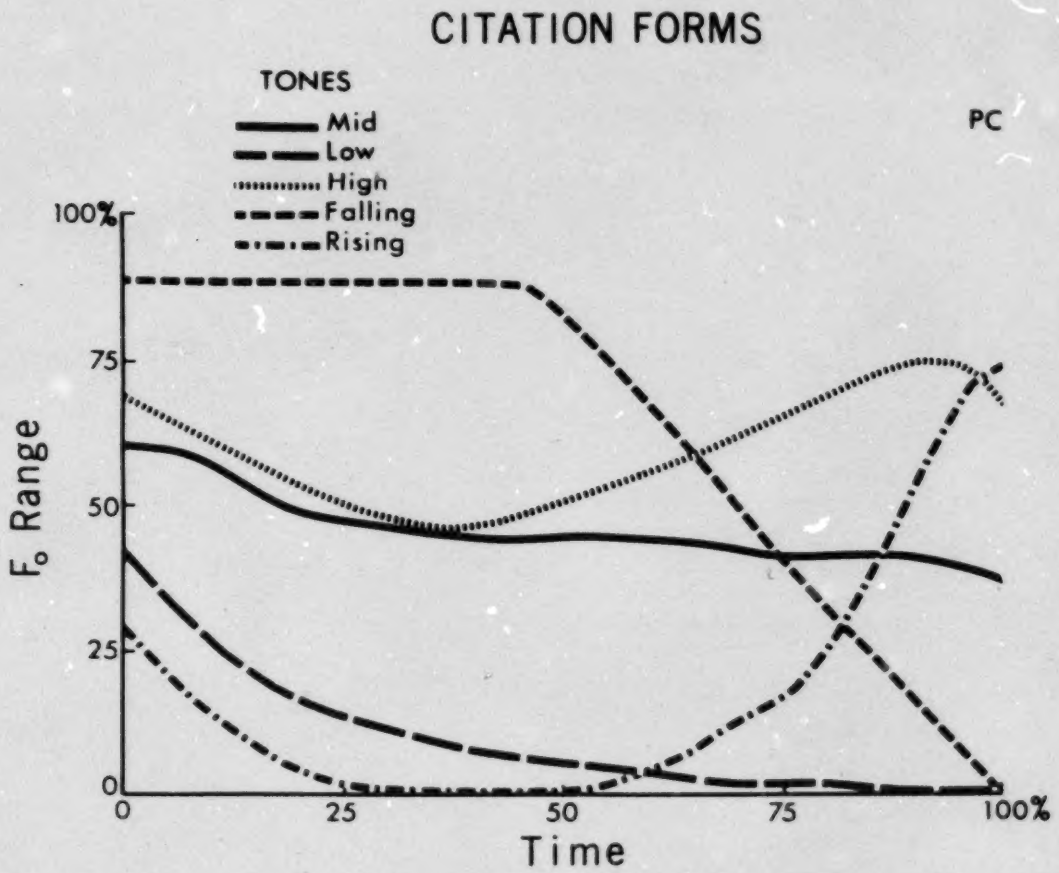


Figure 1: Average F_0 contours on isolated Thai monosyllabic words for one speaker.

by the coded lines, what happens to the context itself. P.C.'s tones in the environment of the following falling tone appear in Figure 3.

The arrays of F_0 curves in the left-hand portions of Figures 2 and 3 clearly support the auditory impression that the tonal system is preserved in these environments. Indeed, inspection of similar displays for the remaining three following contexts, as well as all five preceding contexts for this speaker and the other three speakers, leads to the same conclusion. If we compare the citation forms with the sets of tones in context, the F_0 ranges of the latter are somewhat compressed. Also, the curves are perturbed, especially at their end points, by the preceding and following contexts. Note, for example, that P.C.'s high tone in Figure 2 dips at the end before the following mid tone, while in Figure 3 it stays up before the high beginning of the following falling tone. The falling tone in both figures, coming as it does after the mid tone of the carrier frame, is different from the citation form in that it rises considerably before it falls. In many contexts, as in Figure 2, the high tone and the rising tone might be called high rising and low rising, respectively, as they are almost mirror images of each other with U-shapes that differ mainly in absolute frequency height. It should be noted, however, that the dip of the high

121

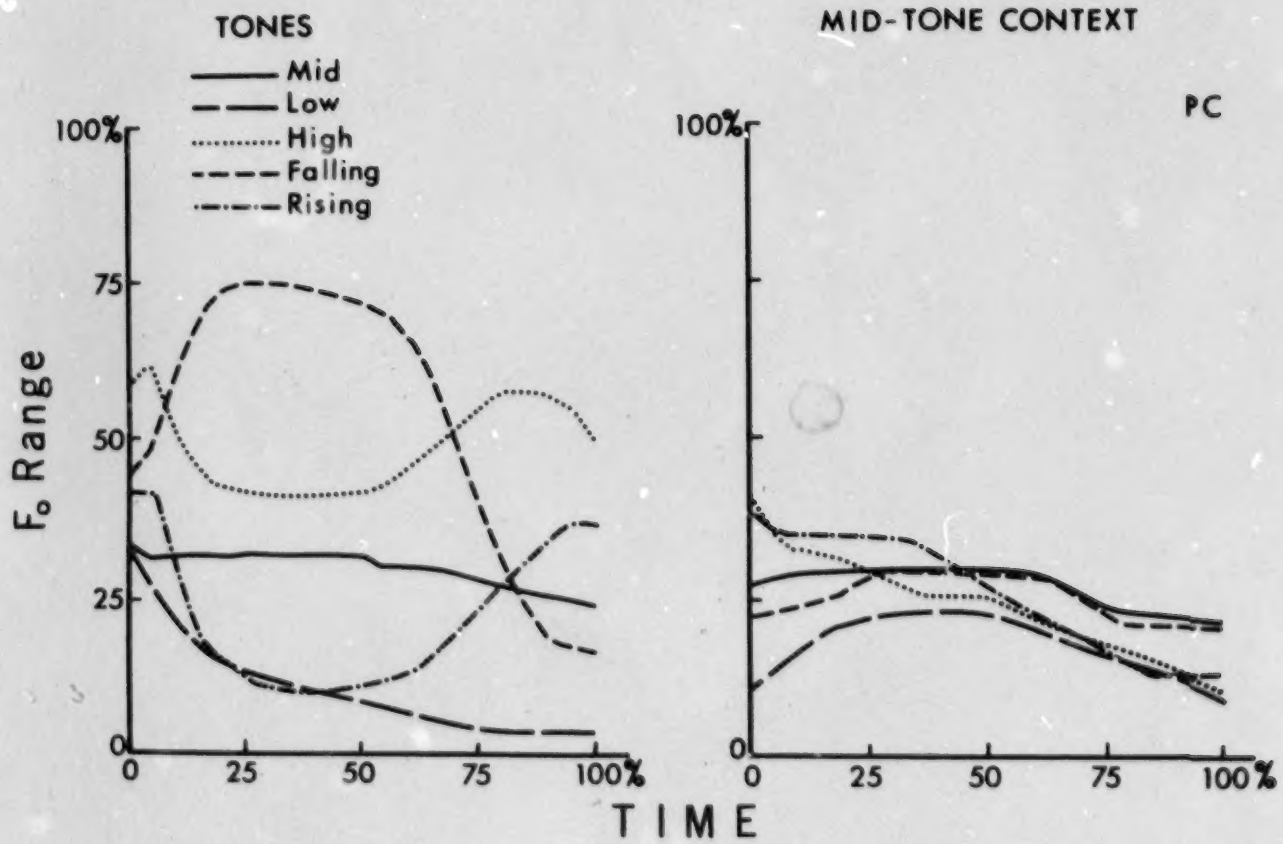


Figure 2: The five tones preceding the mid tone and the resulting variants of the mid-tone context.

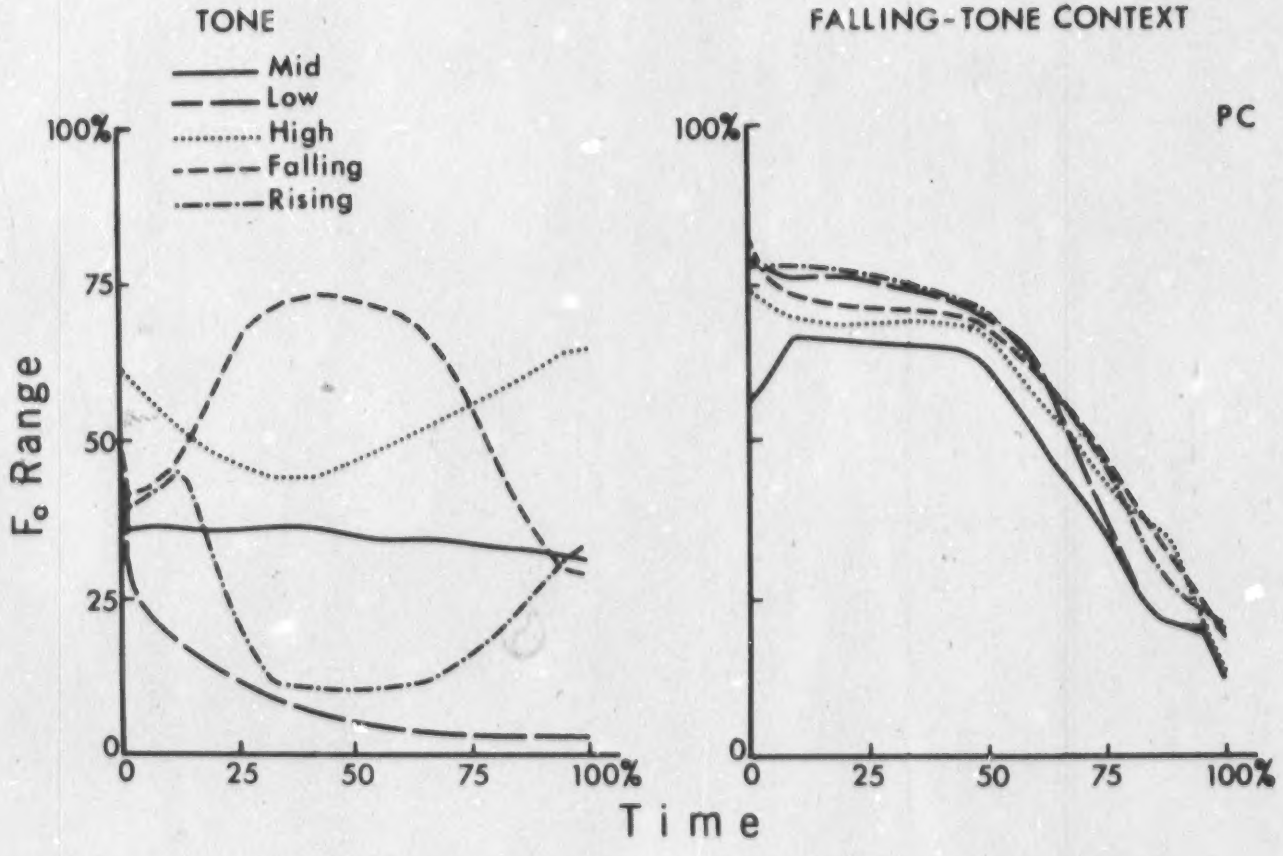


Figure 3: The five tones preceding the falling tone and the resulting variants of the falling-tone context.

123

tone of this type is smaller than that of the rising tone and does not give so obvious an auditory impression of rising as does the rising tone itself.

The families of curves representing the mid-tone and the falling-tone contexts of Figures 2 and 3 form remarkably tight clusters. For all such contexts produced by the four speakers there is very little difficulty in assigning any member of each family of curves to the tone that it is supposed to represent. This is so even though one can see perturbations at the beginning for some of the preceding tones and at the end for some of the following tones.[2] In Figure 2 note the considerable variation in the beginnings of the five variants of the mid tone. In Figure 3 only the variant of the falling tone following the mid tone shows serious initial perturbation, while the others all start in a similar fashion. The other three speakers, however, do not show even this perturbation but rather seem to reset the larynx for a high beginning of the falling tone no matter which of the five tones comes before. They do, nevertheless, show the initial rise of the falling tone after the fixed mid tone of the beginning of the carrier sentence. There may be rhythmic factors at work (Noss, 1972).

DISCUSSION

The data can be understood as reflecting three related effects. (1) The full system of five tones is preserved on monosyllabic Thai words embedded in all possible preceding and following tonal contexts. (2) With citation forms taken as the standard, embedding a tone in running speech causes some perturbation, but generally not enough to damage its identifiability. (3) Laryngeal coarticulation is governed by the specific tonal context; nevertheless, in some contexts coarticulation does not occur for some speakers, and we may suppose that in such instances the larynx receives instructions to reset itself to produce an approximation of the ideal contour for the tone. Even then, rhythmic conditions may prevent the production of the ideal tone as found in citation forms.

In conclusion, it is interesting to note that the data lend no phonetic plausibility to arguments for the specification of the dynamic tones as temporal sequences of features that underlie two of the static tones. The argument, based on purely linguistic reasoning, specifies the falling tone as a sequence of high and low features, and the rising tone as a sequence of low and high features (Leben, 1973; Gandour, 1975). Even the citation forms, let alone the F_0 curves of running speech, provide no acoustic basis for such a claim. It seems psychologically far more reasonable to suppose that the speaker of Thai stores a suitable tonal shape as part of his internal representation of each monosyllabic lexical item.

REFERENCES

Abramson, A. S. (1962) The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments. (Bloomington, Ind.: University Research Center in Anthropology, Folklore, and Linguistics, Publication 20). (IJAL 28.2, Part III, April 1962).

---

[2] In unstressed syllables of compound words certain lexical tones are likely to undergo morphophonemic alternation resulting in the neutralization of some contrasts.

Erickson, D. (1974) Fundamental frequency contours of the tones of standard Thai. Pasaa: Notes and News about Language Teaching and Linguistics in Thailand 4, 1-25.

Gandour, J. (1975) On the representation of tone in Siamese. In Studies in Tai Linguistics in Honor of William J. Gedney, ed. by J. G. Harris and J. R. Chamberlain. (Bangkok: Central Institute of English Language), pp. 170-195.

Han, M. S. and K.-O. Kim. (1974) Phonetic variation of Vietnamese tones in disyllabic utterances. J. Phonetics 2, 223-232.

Henderson, E. (1949) Prosodies in Siamese: A study in synthesis. Asia Major, N.S. 1, 189-215. [Reprinted in Phonetics in Linguistics: A Book of Readings, ed. by W. E. Jones and J. Laver. (London: Longman, 1973), pp. 127-153.]

Leben, W. R. (1973) The role of tone in segmental phonology. In Cononsant Types and Tone, ed. by L. M. Hyman. (Southern California Occasional Papers in Linguistics, No. 1), pp. 115-149.

Lukatela, G. (1973) Pitch determination by adaptive autocorrelation method. Haskins Laboratories Status Report on Speech Research SR-23, 179-208.

MacNeilage, P. F. (1970) Motor control of serial ordering in speech. Psych. Rev. 77, 182-196.

Noss, R. B. (1972) Rhythm in Thai. In Tai Phonetics and Phonology, ed. by J. G. Harris and R. B. Noss. (Bangkok: Central Institute of English Language), pp. 33-42.

Palmer, A. (1969) Thai tone variants and the language teacher. Language Learning 19, 287-299.

Thai Tones as a Reference System*

Arthur S. Abramson+

ABSTRACT

Five monosyllabic words of Central Thai were recorded by ten native speakers and randomized into composite and individual tests. In the composite tests the utterances of all the speakers were randomized to prevent adaptation to any one speaker. In each of the individual tests, only one speaker was used. Identifications were provided by 34 native speakers. The composite tests yielded a fair amount of confusion between the mid and low tones. This confusion was virtually eliminated in the individual tests. We conclude that tones free of a linguistic context are better identified when the listener has access to the speaker's tone space. This effect is best shown by the mid tone, which is likely to be confused with the low tone. The other four tones, which have much more F_0 movement and concomitant variation in amplitude, preserve their perceptual integrity more easily.

An important question about tone languages is how much information is needed by the listener to identify words minimally differentiated by phonologically relevant tones. Relative rather than absolute values of the fundamental frequency (F_0) of the voice are normally found to provide the major acoustic cues for the identification of phonemic tones even in the presence of other phonetic features.[1] Certain tones of a given system may be quite identifiable in isolated

---

*An oral version of this paper was delivered at the 5th Essex Symposium on Phonetics, University of Essex, Colchester, England, 25-27 August. It is to appear in Studies in Tai Linguistics in Honor of Fang-Kuei Li, ed. by T. W. Gething (Bangkok: Central Institute of English Language, in press).

+Also University of Connecticut, Storrs.

Acknowledgment: While most of the analysis of data was performed at Haskins Laboratories, the data themselves were collected while the author was on sabbatical leave in Thailand on research fellowships from the American Council of Learned Societies and the Ford Foundation Southeast Asia Fellowship Program. I gratefully acknowledge the hospitality of Dr. Udom Warotamasikkadit, Dean, the Faculty of Humanities, Ramkhamhaeng University, and Mrs. Mayuri Sukwiwat, Director, the Central Institute of English Language, both in Bangkok.

---

[1] For example, creaky voice, voice breaking, and amplitude variations.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

citation forms of syllables, while others may have to be embedded in a linguistic context for easy recognition. Even the latter tones, however, may enjoy high identifiability in isolation if the listener is free to adapt his perception to the tone system of a given speaker. The perceptual assessment of the tones of a language may well be analogous to that of a vowel space. In the present study, experiments were designed to test the hypothesis that in one tone language, Central Thai (Siamese), listeners would do less well in identifying the tones when deprived of an opportunity to become oriented to the tonal space of the speaker.

Thai is traditionally viewed as having five phonemic tones.[2] The maximum five-way differentiation can appear only on syllables ending in diphthongs, long vowels, or nasal consonants.[3] In principle then, such a syllable can manifest itself as five different monosyllabic words or morphemes, each with its own tone. Every item in the Thai lexicon is characterized by phonologically relevant tone as well as by consonantal and vocalic features. The five tones can be divided into two groups, the dynamic tones and the static tones (Abramson, 1962:9). In this scheme, the rather sharp downward F_0 movement of the falling tone and upward movement of the rising tone place them in the dynamic category. It is their movements rather than their endpoints that seem to characterize them. Since the high, mid, and low tones have sounded to many observers as if they simply occupy three pitch levels, they are classified as static; nevertheless, these tones too, especially the high and low, also show some movement, although it tends to be fairly slow (Abramson, 1962:120-127; Erickson, 1974).[4] This will be relevant to the discussion of the results of the experiments to be presented here. If the static/dynamic dichotomy is perceptually valid, we might expect the static tones to be more readily confused under difficult listening conditions than the dynamic tones.

PROCEDURE

The following Thai monosyllabic words minimally differentiated by tone were recorded by ten native speakers of Central Thai, five men and five women:

Tone	Word	Gloss
Mid	/khaa/	'a grass (Imperata cylindrica)'
Low	/khàa/	'galangal, a rhizome'
Falling	/khâa/	'slave, servant'
High	/khá́a/	'to engage in trade'
Rising	/khǎa/	'leg'

[2] Here I shall not enter into the question of whether the five tones can be profitably analyzed into some smaller number of underlying distinctive features. This question has recently been discussed by Gandour (1975).

[3] Other syllable types support fewer than five tones.

[4] Without undue strain, one can hear pitch glides in the high and low tones. Indeed, F_0 contours of the five tones in running speech (Abramson, in press) render the dynamic/static dichotomy unclear; however, it is mentioned here because of its apparent auditory usefulness.

These tape recordings were randomized into composite tests and individual tests. In the composite tests the ten speakers and all their productions were randomized to make it impossible for the listener to predict at each moment not only what word was going to be said but also which of the ten voices was going to say it. In each of the individual tests the speaker remained the same throughout; the only uncertainty for the listener was the identity of each word as it was played into his headphones. Responses written in Thai script were provided by 34 native speakers over a period of approximately one month. The composite test, in two randomizations, was administered first. In each of the test orders there were three tokens of each word for each of the ten speakers. Each of the ten individual tests was prepared in four randomizations with five tokens of each test order. Since there was not enough time to administer more than one test order for each individual test, care was taken to ensure that no single randomization was presented to the subjects more often than once a week.[5] All the recordings were checked for acceptability by the speakers themselves and the experimenter.

RESULTS

The responses of all 34 subjects to the composite tests are displayed in percentages in the form of a confusion matrix in Table 1. The stimuli are listed in the first column of the matrix and the response labels are arranged across the top. Correct responses to the stimuli as intended are entered in the cells along the diagonal from the upper-left-hand side to the lower-right.[6] An overall percent correct of 94.4 seems rather high until one observes a concentration of errors in one sector of the confusion matrix, nearly all caused by confusion between only two of the tones, the mid and low. The mid tone is heard

TABLE 1: Composite tests. Pooled data for 10 speakers.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	82.4	17.0	0.3	0.3	
	Low	6.9	92.9	0.1	0.1	0.1
	Falling	0.2	0.7	98.9		0.2
	High	0.3	0.1	0.3	98.7	0.6
	Rising	0.1		0.3	0.1	99.6

*Responses = 8847; subjects = 34; percent correct = 94.4.

[5] The tests were run in the language laboratory of Ramkhamhaeng University, Bangkok, Thailand. The subjects, all undergraduates, were chosen and instructed by Miss Phaanit Chotibut of the Department of English; her enthusiastic help is much appreciated.

[6] Variations in the number of responses and the number of subjects in the tables to follow indicate fluctuations in attendance at the test sessions and occassional omissions on the answer sheets.

as low 17 percent of the time, and the low tone is heard as mid 6.9 percent of the time. These confusions are largely caused by four of the speakers whose data will be displayed separately. In addition, about half of the test subjects are responsible for most of the confusions. The responses to the ten separate individual tests are pooled in Table 2. We see an overall improvement of 5 percent, but what is more important is the virtual elimination of the confusions between the mid and low tones. There is also a smaller scattering of errors over the rest of the matrix.

TABLE 2: Individual tests. Pooled data for 10 speakers.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	99.2	0.7		0.1	
	Low	0.9	99.0		0.1	
	Falling		0.1	99.5	0.2	0.1
	High	0.1		0.2	99.5	0.1
	Rising	0.1		0.2		99.7

*Responses = 7450; subjects = 34; percent correct = 99.4.

The major results of the experiment are seen more clearly if we examine the responses to the productions of the four people responsible for most of the errors. The selection criterion was a score of 75 percent or less in any cel of a speaker's confusion matrix for the composite test. This threshold was based on the following reasoning. The chance level for each stimulus is nominally 20 percent, since there was a five-way choice; however, the single serious confusion, that between the low and mid tones, yields, in fact, a chance level of 50 percent. The threshold chosen was halfway between 50 and 100 percent; thus, three speakers had a score of 75 percent or less for the mid tone and one speaker for the low tone in the composite test.

In Tables 3 and 4 we see the composite and individual data, respectively, for Speaker A, a woman. Table 4 shows an overall improvement of 7.4 percent, mostly to be attributed to the virtual elimination of a third of the responses to the mid tone as low in the composite test. In addition, the 3.4 percent identification of high as rising also disappears.

The data for Speaker C, a man, appear in Tables 5 and 6. The effect is even more striking here in that more than half of the responses to the mid tone in the composite test are called low, while there are no responses to it as low in the individual test. It is true, however, that the low tone itself drops a few percentage points in the individual test with the errors assigned to the mid tone; that is, this small confusion in the composite test is not removed in the individual test.

The data for Speaker F, a man, are shown in Tables 7 and 8. They conform to the general trend but in a more satisfyingly clearcut manner, since all scores become 100 percent along the diagonal in Table 8.

TABLE 3: Speaker A's composite data.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	66.1	33.3		0.6	
	Low	0.6	99.4			
	Falling		0.6	99.4		
	High			0.6	96.0	3.4
	Rising				1.1	98.9

*Responses = 885; subjects = 34; percent correct = 92.1.

TABLE 4: Speaker A's individual data.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	98.0	2.0			
	Low	0.7	99.3			
	Falling			100		
	High				100	
	Rising					100

*Responses = 750; subjects = 30; percent correct = 99.5.

TABLE 5: Speaker C's composite data.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	39.0	58.7	2.3		
	Low	3.4	96.6			
	Falling		0.6	99.4		
	High	0.6			99.4	
	Rising					100

*Responses = 885; subjects = 34; percent correct = 90.9.

TABLE 6: Speaker C's individual data.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	99.2			0.8	
	Low	5.4	93.9		0.8	
	Falling		1.5	97.7	0.8	
	High	1.5			97.7	0.8
	Rising		0.8	0.8		97.7

*Responses = 650; subjects = 26; percent correct = 97.4.

TABLE 7: Speaker F's composite data.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	75.1	24.9			
	Low	2.3	97.7			
	Falling	0.6	2.3	96.6		0.6
	High			2.3	97.7	
	Rising					100

*Responses = 885; subjects = 34; percent correct = 93.5.

TABLE 8: Speaker F's individual data.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	100				
	Low		100			
	Falling			100		
	High				100	
	Rising					100

*Responses = 750; subjects = 30; percent correct = 100.

The data for Speaker J, a man, shown in Tables 9 and 10, support the hypothesis stated at the outset but do so in a different way. It is not the mid tone that tends to be misheard in the composite test but rather the low tone, which is heard as mid 35.4 percent of the time.

TABLE 9: Speaker J's composite data.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	100				
	Low	35.4	64.6			
	Falling	0.6		98.9		0.6
	High				100	
	Rising	0.6				99.4

*Responses = 885; subjects = 34; percent correct = 92.7.

TABLE 10: Speaker J's individual data.*

		Percent Responses				
	Labels:	Mid	Low	Falling	High	Rising
Stimuli	Mid	100				
	Low	0.7	99.4			
	Falling			100		
	High				100	
	Rising					100

*Responses = 775; subjects = 31; percent correct = 99.9.

For the other six speakers the general pattern is the same, although the effects are somewhat smaller. The mid tones of five of them are sometimes heard as low in the composite test while the low tones of the sixth person are sometimes heard as mid. These effects disappear in the individual tests.

The 34 listeners were not equally confused by the utterances of the mid and low tones produced by the four speakers of Tables 3-10. Every subject made at least some errors, but some listeners were much more affected than others. In addition, the subjects varied somewhat as to which speaker confused them the most. Every one of the four speakers had 100 percent correct responses to the two tones in question for at least a few listeners. For example, Speaker A caused no errors in the responses of eight subjects to the composite test. Even Speaker C, whose mid tone was heard as low 58.7 percent of the time, had three listeners who made no errors in labeling his productions of the two tones.

DISCUSSION

We may conclude that in Thai, and probably in other tone languages, phonemic tones free of a linguistic context are better identified when the listener has access to the speaker's tone space. The two Thai tones that are most vulnerable to confusion are the mid and low tones, which are typically characterized by very little movement of the fundamental frequency of the voice over time. This can be seen in Figure 1 for one of the speakers least productive of errors in this study; that is, not one of the four singled out for special attention. The figure shows Speaker P.C.'s F_0 curves normalized for time and stated as percentages of the total voice range used by her for the test items. The falling, rising, and high tones--thus two dynamic tones and one static tone--appear to have such distinctive contours that their great invulnerability to confusion in the composite condition is not at all surprising. Indeed, the high tone of citation forms typically includes quite audible glottal constriction or creak.

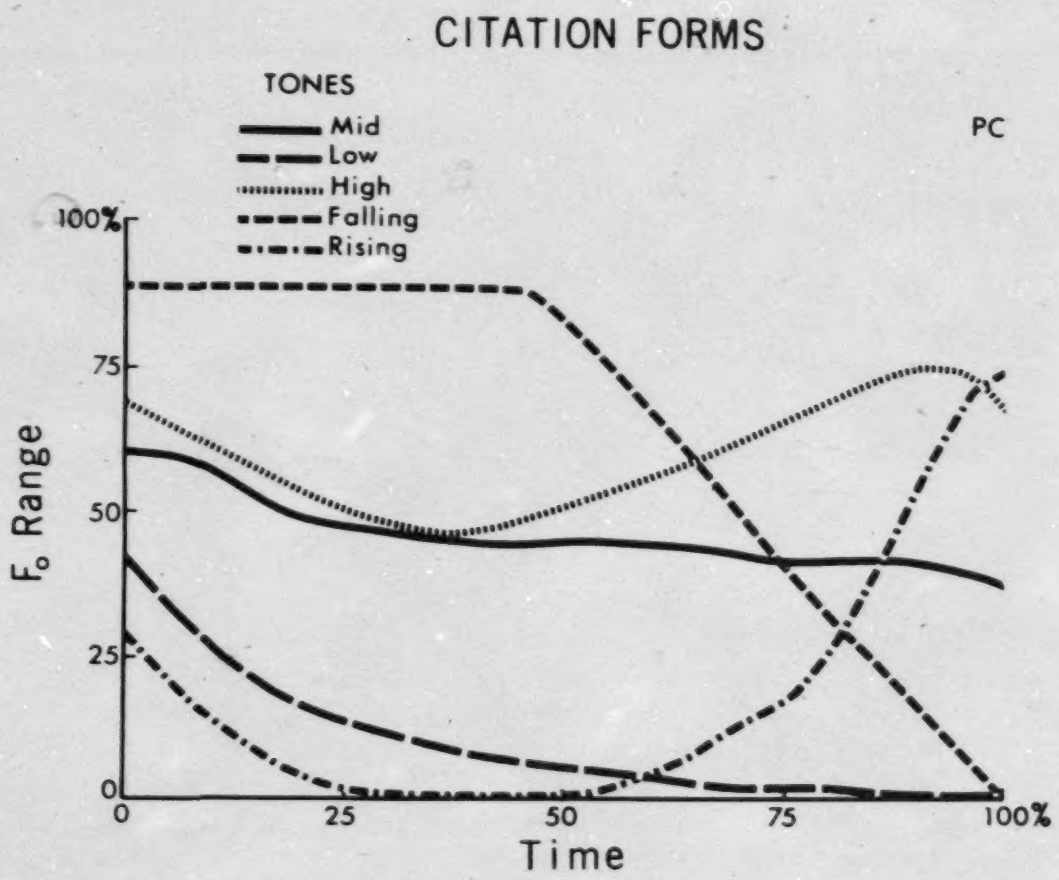


Figure 1: Average F_0 contours of Speaker P.C. produced for the tests.

The question still arises as to why it is that the mid tone is normally confused with the low tone rather than the opposite in the composite condition.[7] Figure 1 shows that the low tone, as compared with the mid tone, drops fairly abruptly and then descends slowly toward the bottom of the voice range. When this characteristic movement is very evident to the listener, it may render the low tone robust. Such a contour would seem to be a sufficient cue for the low tone but not a necessary one, because the down drift of the mid tone in the confusion of the composite condition may be enough to make some listeners uncertain and cause them to assign it to the only possible other choice, namely, the low tone.

The opposite effect for Speaker J (Table 9), labeling the intended low tone as mid, is more difficult to explain. It should be noted that when a speaker's low tone reaches the bottom of his voice range quite early and stays there until the end, vocal fry can often be heard in citation forms. The low end of Speaker J's voice range is defined by the bottom point of his rising tone and not by his low tone. (For Speaker P.C. in Figure 1 the bottom of her range is reached by both these tones as well as the falling tone.) We may speculate then that Speaker J's mid-tone contour is robust because its shape cannot be mistaken for anything else even in the composite condition, while his low tone simply does not possess the voice quality often associated with that tone when it clings for a good part of its duration to the bottom of the voice range.

Given the disagreements in the literature over the stability of this tonal opposition (Noss, 1954, 1964; Abramson, 1962, 1972, 1975), it is interesting to find that this contrast is so vulnerable to confusion in the composite condition. Noss (1975:279) comments, "It would appear that, while individual speakers keep the mid- and low-tone contours distinct, there may be little or no differentiation of these contours between speakers." His view, although somewhat overstated, receives support from a comparison of the composite and individual conditions of the present study.

Finally, the individual differences observed in production and perception lead to one other generalization. Some Thai speakers seem to provide minimal cues for the distinction between the mid and low tones. Many listeners require more than these minimal cues if they are to identify the tones of those speakers when no opportunity is given to become accommodated to the tonal space of each such speaker.

REFERENCES

Abramson, A. S. (1962) The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments. (Bloomington, Ind.: University Research Center in Anthropology, Folklore, and Linguistics, Publication 20). (IJAL 28.2, Part III, April 1962).

Abramson, A. S. (1972) Tonal experiments with whispered Thai. In Papers in Linguistics and Phonetics to the Memory of Pierre Delattre, ed. by A. Valdman. (The Hague: Mouton), pp. 31-44.

Abramson, A. S. (1975) The tones of Central Thai: Some perceptual experiments. In Studies in Tai Linguistics in Honor of William J. Gedney, ed. by J. G. Harris and J. R. Chamberlain. (Bangkok: Central Institute of English Language), pp. 1-16.

[7] The data do not support any influence of sex.

Abramson, A. S. (in press) The coarticulation of tones: An acoustic study of Thai. In Proceedings of the 8th International Congress of Phonetic Sciences. [Also in Haskins Laboratories Status Report on Speech Research SR-44 (this issue).]

Erickson, D. (1974) Fundamental frequency contours of the tones of standard Thai. Pasaa: Notes and News about Language Teaching and Linguistics in Thailand 4, 1-25.

Gandour, J. (1975) On the representation of tone in Siamese. In Studies in Tai Linguistics in Honor of William J. Gedney, ed. by J. G. Harris and J. R. Chamberlain. (Bangkok: Central Institute of English Language), pp. 170-195.

Noss, R. B. (1954) An outline of Siamese grammar. Unpublished Ph.D. dissertation, Yale University.

Noss, R. B. (1964) Thai Reference Grammar. (Washington, D.C.: Foreign Service Institute).

Noss, R. B. (1975) How useful are citation forms in synchronic Thai phonology? In Studies in Tai Linguistics in Honor of William J. Gedney, ed. by J. G. Harris and J. R. Chamberlain. (Bangkok: Central Institute of English Language), pp. 274-284.

Some Electromyographic Measures of Coarticulation in VCV Utterances*

Thomas Gay+

ABSTRACT

The purpose of this experiment was to study the motor patterns that underlie articulator movements during the production of certain vowel-consonant-vowel (VCV) syllables. Electromyographic (EMG) data were attained from the genioglossus and orbicularis oris muscles of two subjects. The speech material contained both VCV and CVVC contrasts. The data showed surprisingly little in the way of carryover effects of the first vowel on the EMG signals of the second vowel: peak amplitudes of the same second vowel were little changed for different first vowels. For all utterances where the genioglossus muscle was active for both the first and second vowels, a trough appeared in the EMG envelope during the time of consonant production. Both findings were interpreted as supporting the existence of a neutral tongue body position during consonant production.

In a recent cinefluorographic study of vowel production (Gay, 1974), we showed, that in a vowel-consonant-vowel (VCV) utterance, vowel targets are reached efficiently and with a high degree of precision, irrespective of changes in the intervocalic consonant or in the vowel that either precedes or follows the consonant. This high degree of articulatory accuracy (evident especially for /i/ and /u/, and only slightly less so for /a/) supports a spatial target or target field specification of vowels (MacNeilage, 1970).

A basic assumption of a target-based model of speech production is that the underlying motor input to the target-directed movement is characterized by what MacNeilage (1970:184) calls, "an elegantly controlled variability of response to the demand for a relatively constant end." In other words, depending on the particular vocal-tract shape or articulator position for a preceding phone, movement toward a given target will be controlled by any of a number of different motor strategies.

---

*This paper was delivered at the 5th Essex Symposium on Phonetics, University of Essex, Colchester, England, 25-27 August 1975.

+Also University of Connecticut Health Center, Farmington.

Acknowledgment: This research was supported in part by a grant from the National Institute of Neurological Diseases and Stroke (NS-10424), the National Science Fonndation (GSOC 740 3725).

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

In order to determine whether such different strategies do, indeed, exist for target-directed movements that originate from different contexts, we undertook a corresponding electromyographic (EMG) study of coarticulation in VCV utterances. In this experiment, we used EMG to study the motor patterns that underlie the vowel gesture in VCV sequences using the same subjects and speech material as in our original X-ray motion picture experiment. A second purpose of this experiment was to test Öhman's (1966) hypothesis that articulator movement in a VCV sequence is essentially diphthongal, with the consonant gesture superimposed on the basic vowel-to-vowel movement.

METHOD

Subjects were the same two native speakers of American English that we studied in our earlier X-ray experiment (FSC and TG). The speech material contained both VCV and CVVC contrasts. The VCVs consisted of the consonants /p,t,k/ and the vowels /i,a,u/ in a trisyllable nonsense word of the form, /kV_1CV_2pə/, where V_1 and V_2 were all possible combinations of /i,a,u/ and C was either /p/, /t/, or /k/. Corresponding CVVC syllables, of the form /kV_1V_2pə/, were included for purposes of comparing motor patterns between two vowel sequences when one set of utterances was separated by a consonant and the other was not. This contrast was used to test Öhman's (1966) vowel-to-vowel movement hypothesis. Examples of the two types of utterances are /kipapə/ - /kiapə/ and /kutipə/ - /kuipə/. Both sets of utterance types were randomized into four lists, each of which was read five times by the two speakers at normal speaking rates. The carrier phrase, "It's a ..." preceded each utterance, and syllable stress was on the second vowel.

For both subjects, EMG recordings were obtained from the genioglossus and orbicularis oris (superior) muscles. The genioglossus muscle makes up the bulk of the tongue and acts in bunching and protruding the tongue during the production of both /i/ and /u/. The orbicularis oris is active in both closing and rounding the lips. Following usual practice, conventional hooked-wire electrodes were used for the insertions, and the data were recorded on magnetic tape and later processed using the Haskins Laboratories EMG data processing system.

RESULTS

The first question we will address is whether different motor strategies are employed in the control of target-directed movements for a vowel as a function of differences in both the preceding consonant and the preceding vowel ahead of the consonant.

Figures 1 and 2 summarize the effects of the consonant on the activity level of the genioglossus muscle for the vowel /i/. These figures show the averaged EMG signals for the utterances /kupipə/, /kutipə/, and /kukipə/ for both subjects. Each plot contains three peaks. The first represents the muscle activity for the initial /k/, the second for the first vowel (which is merged with the peak for the intervocalic /k/ in /kikipə/), and the third for the second vowel. Levels for the initial /k/ vary somewhat as do those for the first vowel. We attribute this variability to the fact that the first vowel was destressed, and probably variably so, relative to the second vowel; such stress allophones have been shown to exist (Harris, 1973). Note, however, that for both subjects the EMG signals for the second vowel are surprisingly similar in peak height. The range of variability of peak height for the second vowel was,

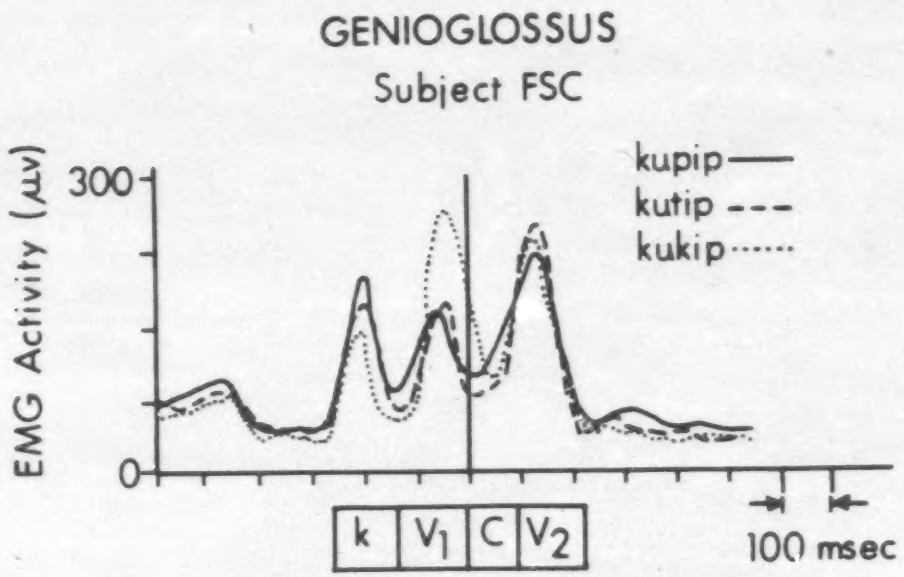


Figure 1: Averaged EMG activity showing the effect of the intervocalic consonant on the production of the vowel, Subject FSC.

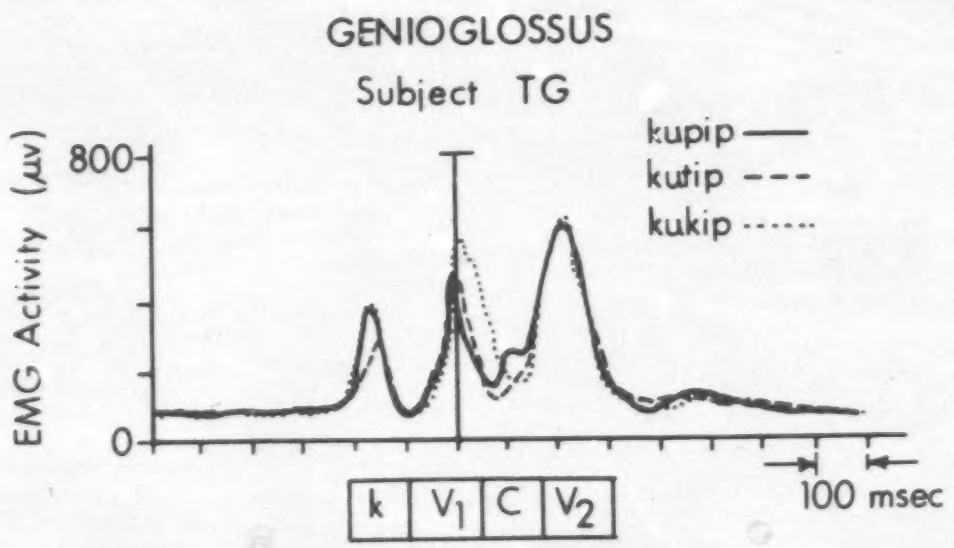


Figure 2: Averaged EMG activity showing the effect of the intervocalic consonant on the production of the vowel, Subject TG.

except in one instance, within 25 μv. This is a particularly narrow range considering both the place differences of the three consonants and the fact that the peak signal strength approaches 300 μv for one subject and 800 μv for the other.

Carryover effects of the first vowel (ahead of the intervocalic consonant) on the muscle signals for the second vowel are similarly absent. This is illustrated in Figures 3 and 4. Here, what we interpret as stress effects again appear for the initial /k/. Likewise, the unexpected absence of any effect of different first vowels on the muscle signal for the second vowel is also evident. Again, too, the absence of carryover effects is consistent for both subjects and for all but one utterance, and peak heights occur within a range of 25 μv.

The results described above show very small or no carryover coarticulation. Somewhat greater effects have been shown for VCV utterances by Bell-Berti and Harris (1975). Their utterances contained the same vowels (/i,a,u/) as ours and, in addition, were balanced for stress. They showed almost no anticipatory coarticulation effects of the second vowel on the first vowel, but somewhat greater carryover coarticulation than in the present study. A possible explanation is that the Bell-Berti and Harris syllables were spoken at a considerably faster rate than ours; perhaps fast speech effects were reflected by greater degrees of motor variability.

The extreme interpretation of these results would be the resurrection of the notion of motor invariance in the production of a speech gesture. Our more moderate interpretation, however, is that the observed invariance in the EMG signals for the vowel reflects rather an invariant or neutral tongue body target for the consonant. In other words, the signals are the same because the pattern of movement from consonant target to vowel target are the same.

Further evidence of a tongue body target during the production of a labial or alveolar consonant appears in several forms. Referring back to Figures 1 and 2, we can see that during consonant production, a trough appears in the genioglossus curve. The presence of this trough represents a cessation of genioglossus muscle activity during the production of the consonant, and hence, what we interpret as a break in continuous movement from the first vowel to the second. Figure 5 shows the EMG data for the corresponding CVVC utterance /kuip/. Although a small dip appears in the envelope, the basic pattern is one of uninterrupted activity throughout the entire vowel-to-vowel sequence. The interruption in EMG activity in the VCV, when compared to the continuous activity in the CVVC, would seem to argue against Öhman's (1966) hypothesis that vowel-to-vowel movement in a VCV is basically diphthongal. A more convincing "trough" illustration, however, is in the VCV utterance where both first and second vowels are the same. Figures 6 and 7 illustrate these effects for the utterances /kipipə/ and /kitipə/ for both subjects. Here, the basic pattern is the same as before: a peak for the first vowel followed by a trough for the consonant and another peak for the second vowel. Again, our reasoning is that if the tongue body did not assume a different position for the consonant, the EMG envelope would contain one broad peak representing positional constancy for /i/, rather than two peaks separated by a distinct trough. This is consonant with the report of Bell-Berti and Harris (1974).

A trough in muscle activity during the consonant also appears in the lip-rounding component of the vowel. Figure 8 shows both the genioglossus and

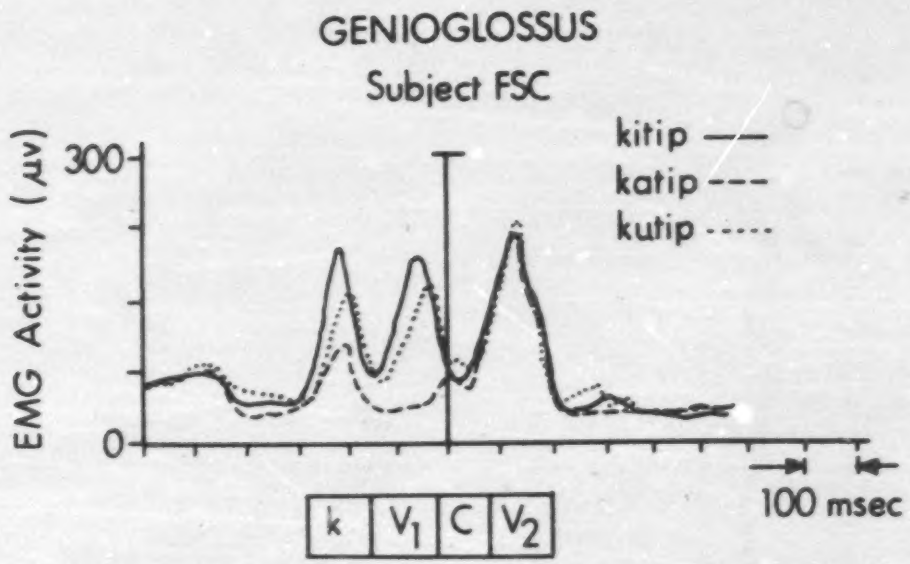


Figure 3: Averaged EMG activity showing the effect of the first vowel on the production of the second vowel, Subject FSC.

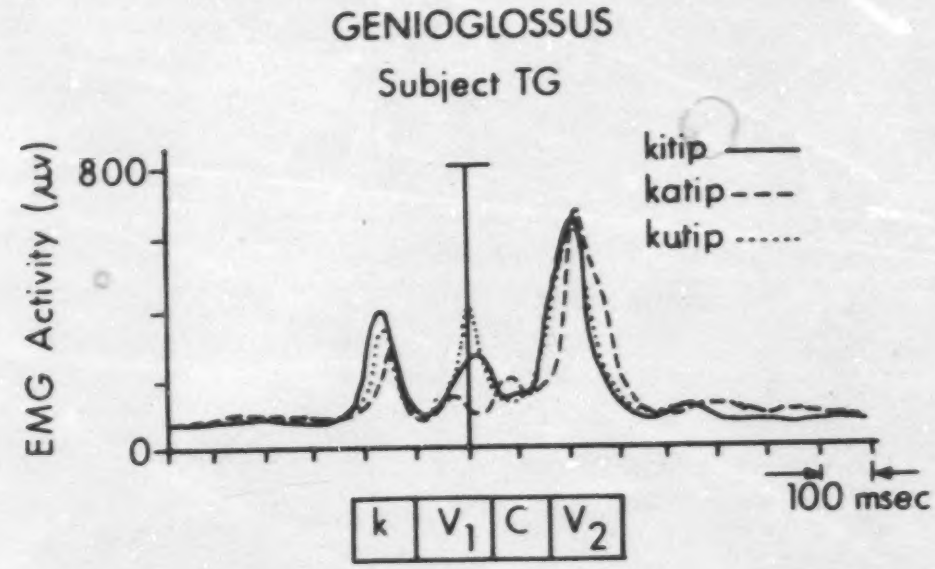


Figure 4: Averaged EMG activity showing the effect of the first vowel on the production of the second vowel, Subject TG.

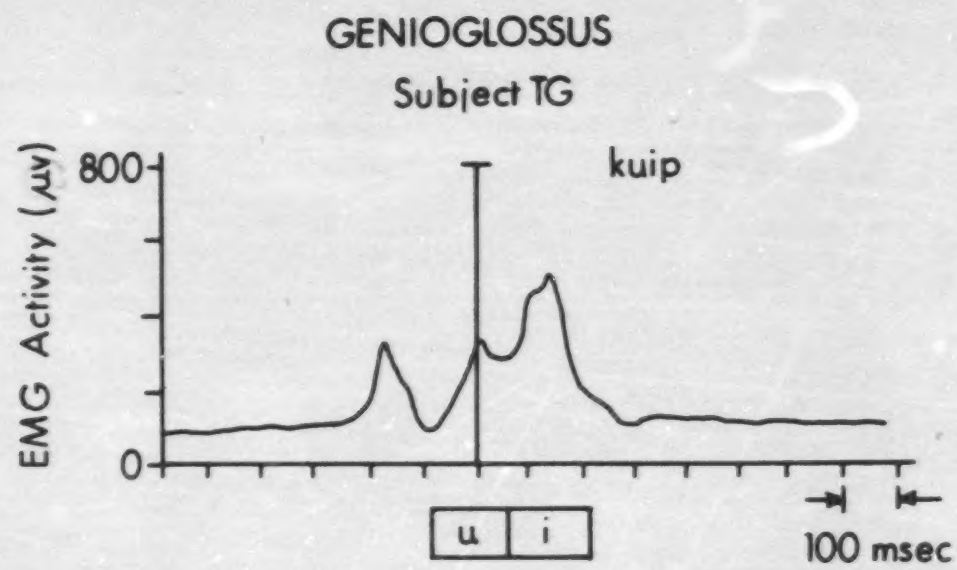


Figure 5: Averaged EMG activity for the CVVC sequence, /kuipə/, Subject TG.

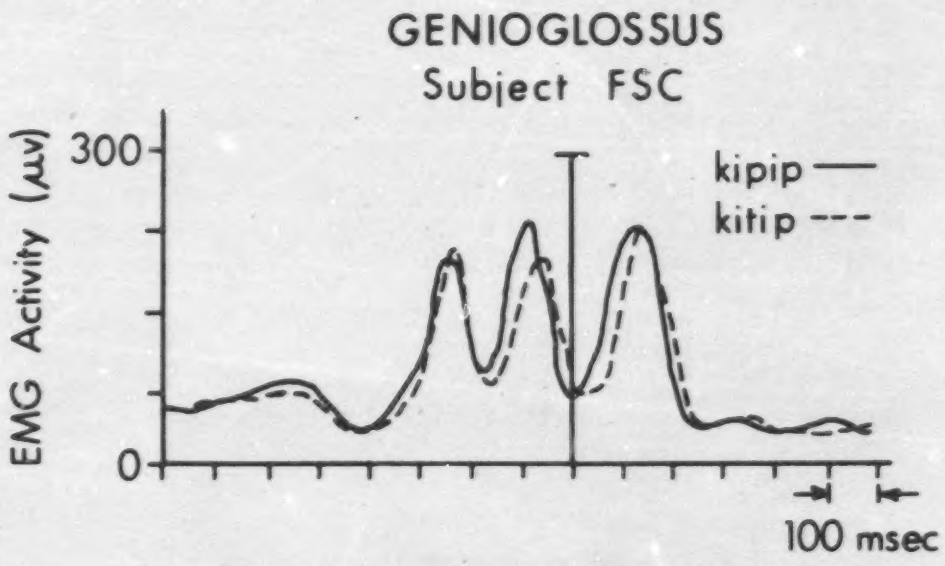


Figure 6: Averaged EMG activity showing two separate vowel peaks for a VCV when the first and second vowels are the same, Subject FSC.

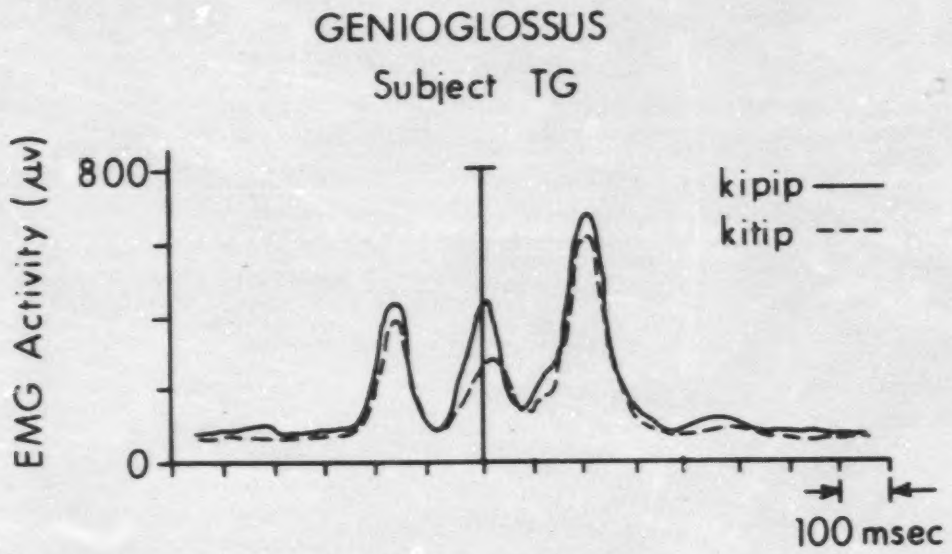


Figure 7: Averaged EMG activity showing two separate vowel peaks for a VCV when the first and second vowels are the same, Subject TG.

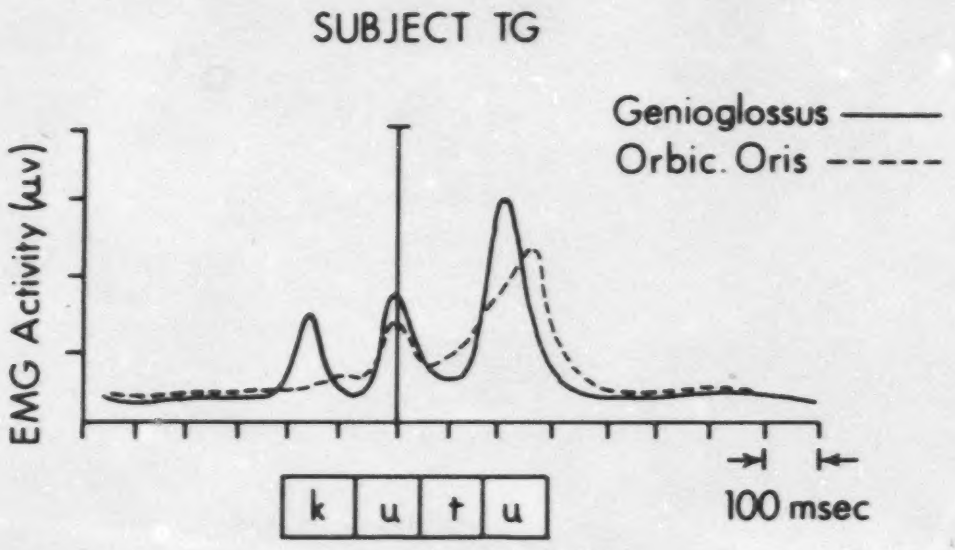


Figure 8: Averaged EMG activity for both the genioglossus and orbicularis oris muscles during the production of the utterance, /kutupə/.

orbicularis oris muscle activity data for the utterance /kutupə/ for Subject TG. Note how both curves rise for the vowel and then dip for the consonant before rising again for the vowel. The only anticipatory movement seems to be a slight time lead for orbicularis oris activity for the second vowel. The absence of anticipatory lip rounding for the second vowel during the consonant contradicts the X-ray data of Daniloff and Moll (1968), which showed that lip rounding for a vowel can begin as early as three or four phonemes ahead of the vowel. One explanation for the Daniloff and Moll result might be that those phonemes preceding the vowel in their syllables contained lip-rounding components themselves. For example, in the word "construe," it might be argued that both the /s/ and /r/ are rounded consonants. The nonlabial consonants in our utterances (i.e., /t/ and /k/) are probably not marked by such a "rounding" feature.

CONCLUSIONS

The presence of a trough in the EMG envelope for a vowel during the time of consonant production suggests that a tongue body position or target exists for the consonant. The finding that similar motor patterns underlie the production of a vowel followed by different consonants suggests that this tongue body target is a neutral one. Both findings argue against the ubiquity of anticipatory gestures occurring earlier than one segment ahead of a particular phone (Henke, 1966). This, of course, contradicts the well-known data of Daniloff and Moll (1968) that showed the onset of lip rounding for /u/ to begin as early as four or five segments ahead of the vowel. We would like to propose, however, as an alternative interpretation of the Daniloff and Moll finding, that the so-called "early" onset of lip rounding in their data was not in anticipation of a downstream vowel but rather corresponded to a possible rounding component of one or more of the immediately preceding consonants: /n/, /s/, /t/, /r/.

Although our data do not show any evidence of anticipatory movements in VCVs, we do not propose a strict phoneme-by-phoneme specification of segmental ordering. On the contrary, the evidence for an anticipatory coarticulation field is widespread; we are proposing simply a reappraisal of current views on its size.

REFERENCES

Bell-Berti, F. and K. S. Harris. (1974) On the motor organization of speech gestures. Haskins Laboratories Status Report on Speech Research SR-37/38, 73-77.

Bell-Berti, F. and K. S. Harris. (1975) Some aspects of coarticulation. Paper presented at the 8th International Congress of Phonetic Sciences, Leeds, England, 17-23 August.

Daniloff, R. G. and K. Moll. (1968) Coarticulation of lip rounding. J. Speech Hearing Res. 11, 707-721.

Gay, T. (1974) A cinefluorographic study of vowel production. J. Phonetics 2, 255-266.

Harris, K. S. (1973) Stress and syllable duration change. Haskins Laboratories Status Report on Speech Research SR-35/36, 31-38.

Henke, W. (1966) Dynamic articulatory model of speech production using computer simulation. Unpublished Ph.D. thesis, Massachusetts Institute of Technology.

MacNeilage, P. F. (1970) The motor control of serial ordering of speech. Psychol. Rev. 77, 182-196.
Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Am. 39, 151-168.

Durations of Articulator Movements for /s/-Stop Clusters*

Gloria J. Borden[+] and Thomas Gay[++]

ABSTRACT

Cine X-ray data derived by pellet tracking from three subjects indicate that there are certain basic strategies used for initial /sp/, /st/, and /sk/ clusters that reveal a remarkable economy of effort in tongue movement. In addition, the known acoustic shortening of /s/ before /p/ is seen to be a result of early lip closure relative to the tongue closure delayed by its involvement with the /s/.

In this paper we present some findings from an ongoing study of the fricative /s/ as it is produced by normal speakers either as a single consonant or in cluster with other consonants. In general, consonant clusters in speech are interesting to study because the individual events are so closely timed that they presuppose a complex motor program. The long-range purpose of the study is to investigate the relationships among electromyographic (EMG) recordings taken from tongue muscles, movement data recorded in the form of X-ray motion pictures, and acoustic measurements. Data have been collected thus far from three subjects; from two of these subjects we have acoustic and movement information and from one we have simultaneous EMG and cine X-ray data. The focus of this paper will be on a comparative analysis of the clusters /sp/, /st/, and /sk/ achieved by inspection of acoustic information and movement information as recorded on X-ray motion pictures.

---

*This is a combination of oral papers presented by the authors in 1973: "Production of /s/ in Clusters," presented at the 8th International Congress of Phonetic Sciences, Leeds, England, 17-23 August, and "The Relative Durations of Articulator Movements for /s/-Stop Clusters," presented at the 90th meeting of the Acoustical Society of America, San Francisco, 4-7 November.

[+]Also City College, City University of New York.

[++]Also University of Connecticut Health Center, Farmington.

Acknowledgment: This research is sponsored by the National Institute of Dental Research and the National Institute of Neurological Diseases and Stroke. We are grateful to them, to Dr. Seiji Niimi from the University of Tokyo, and to Dr. J. Daniel Subtelny in Rochester.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]

before noting the features common to all of the speakers, a few of the individual variations should be mentioned. Subject 1 produces /s/ with the tongue tip low behind the lower incisors and the blade high to form the tongue-palate constriction, whereas Subjects 2 and 3 elevate the tongue tip to a position posterior to the upper incisors. In all cases, it was the blade of the tongue, not the tip, forming the constriction, but for the tongue-tip-high subjects a more anterior portion of the tongue blade was involved. These two variant articulations of /s/ are common.

It was noted too that Subject 3 moves his tongue front and back more than the other subjects, but for all subjects vertical movement was more extensive than horizontal movement; that is, all subjects move their jaws, lips, and tongue primarily up and down with much less front-back movement.

The instrumentation used for the first subject has been reported by Borden and Gay (1975). The two subjects recorded since then were recorded and analyzed under a newer system. It is the more recent recording and analysis system that we shall describe in this paper.

Figure 1 shows the position of the subjects for the X-ray films. The subject's head was stabilized and lead pellets of 2.5-mm diameter were attached by means of a cyanoacrylate adhesive to the upper and lower lips and to three positions on the tongue--the tip, the blade, and the dorsum--approximately one inch apart, and a reference pellet was attached at the midpoint of the upper central incisors. The X-ray generator delivered 1-msec pulses at 100 kV to a 9-in. image intensifier tube. The film was recorded with a 16-mm camera at a speed of 60 frames per second.

Figure 1: Lateral X-ray motion pictures record movement of lead pellets attached to the tongue, with pellet on incisors used as reference.

The film was analyzed frame by frame. The method used will be described later, but there are certain strategies used for the production of /sp/, /st/, and /sk/ before the low vowel /ɑ/ that should be noted here. The simplest way to view the strategies used by the subjects is to refer to a schematic representation of relative jaw, tongue, and lip movement. The acoustic signal and the movement data are lined up at the moment of /p/ closure.

The subjects had similar strategies for /st/ and /sk/, but one subject differed for /sp/. Therefore, two strategies will be shown for this cluster. The subjects agreed essentially on articulator movement irrespective of the high or low tongue tip difference.

Figure 2 represents the first strategy for producing the utterance /spɑp/: the simple asynchronous opening of the jaw with the tongue and lower lip. The upper lip moves relatively little. Notice that the jaw that carries the tongue is free to lower for /ɑ/ immediately upon lip closure for the preceding /p/.

Subjects 1 and 2, however, used a less straightforward strategy for /spɑp/ in which the tongue lowered before the jaw and lips. In Figure 3 the lineup point for the acoustic and movement data is lip closure for the first /p/. Again, the upper lip is steady as the lower lip is closing for the /p/. Note that as the lower lip is being elevated, the tongue is beginning to lower for /ɑ/. The movements are not only asynchronous but in opposite directions, the lip coming up as the tongue is going down. The difference in strategy between these subjects and the first shown may be attributed to speaking rate, since the first simpler strategy was used with a relatively slow utterance of /spɑp/. Notice here too that the tongue lowers a bit on its own before being carried the rest of the way by the lowering of the jaw.

For /st/ all three subjects used the same strategy, which is abstracted in Figure 4. For /st/, instead of the tongue lowering as a unit, as was the case for /sp/, the tongue tip and the blade rise to form the blade-alveolar ridge occlusion for /t/, but the more posterior portion of the tongue is free to descend early for the low vowel. Note too that the blade is high for the /s/ as the tongue tip is rising to take a more active part in the stop closure. (Subject 1, who produced /s/ with the tongue tip low, uses this basic strategy for /st/, but the tongue tip is kept relatively low for the alveolar stop as well.) If we look at one example of this strategy for /st/, we can see the differential movement of the tongue pellets. Figure 5 illustrates the film analysis system used in this study. It was produced by first projecting the film image onto a writing surface via an overhead mirror system in order to mark pellet positions and frame numbers that are input to a small laboratory computer via a digitizing tablet. The computer measures the x and y coordinate positions relative to the reference pellet position, converts the measurements to millimeter distances, and stores the data on digital magnetic tape. A second program draws axes and plots the measured data on a display scope from which a hard copy is made.

The horizontal line represents the position of the reference pellet that was placed between the upper central incisors. The vertical line represents the lip closure for the /p/. The utterance here is /stɑpə/. The pellet on the tip of the tongue represented by a square can be seen to elevate slowly for the fricative /s/, rise even more for the alveolar stop /t/, and rapidly descend for the low vowel /ɑ/. The blade pellet (the circles) also remains high for the /st/ and descends upon release of the stop. The more posteriorally placed tongue

/spɑp/

Strategy 1: Synchronous opening of jaw with tongue and lower lip.

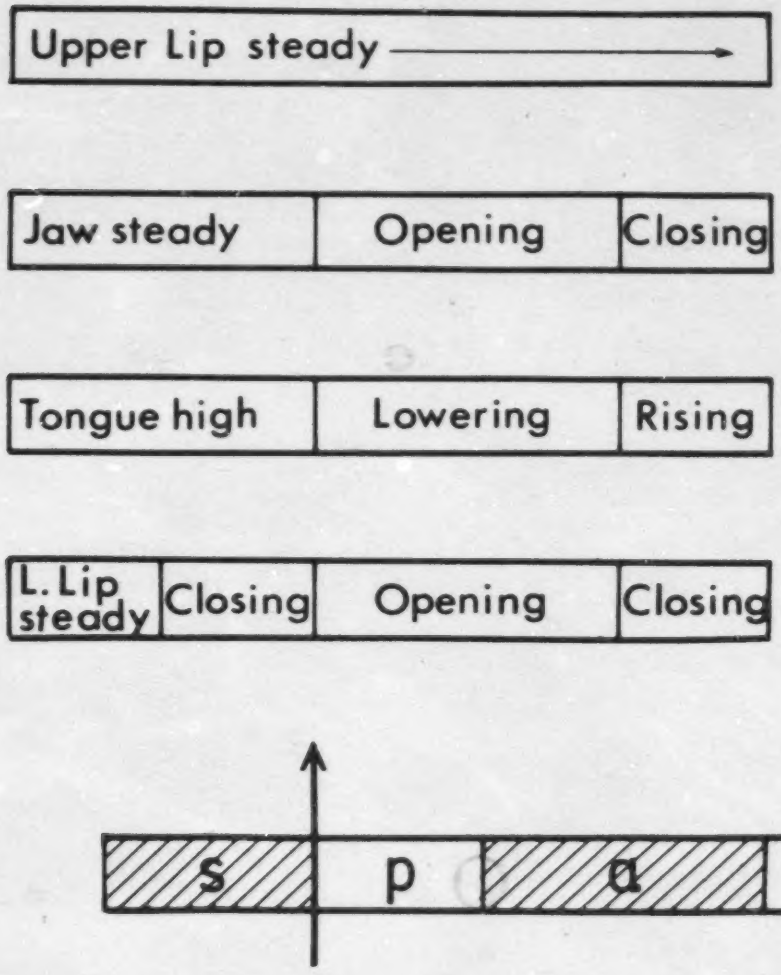


Figure 2: Schematic of relative movement of articulators for /spɑp/. Jaw lowers for /ɑ/ immediately upon lip closure for /p/.

/spɑp/

Strategy 2: Tongue lowers before jaw and lips.

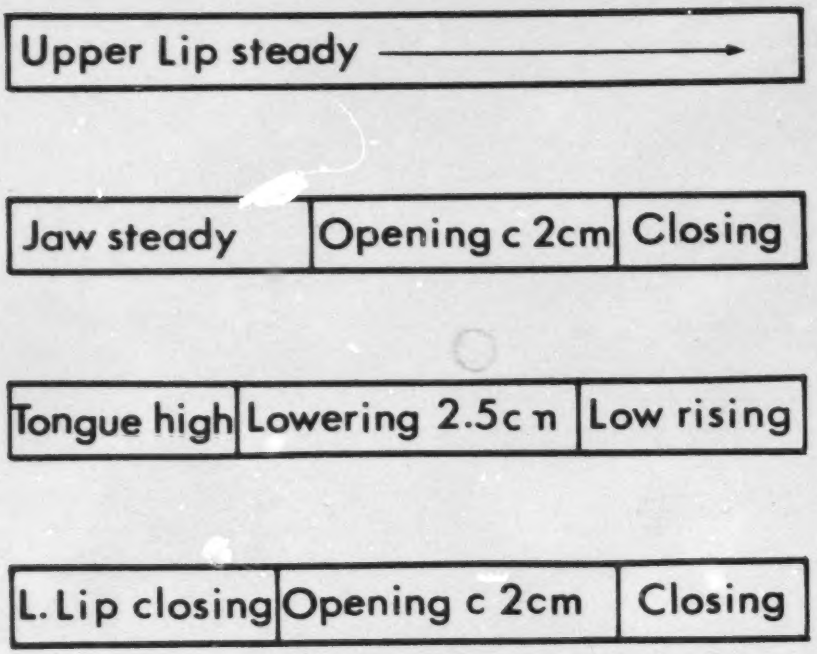


Figure 3: Schematic of asynchronous movements of articulators for /spɑp/. Tongue lowers as a unit.

/stɑpə/

Strategy: Back of tongue descends early for /ɑ/.

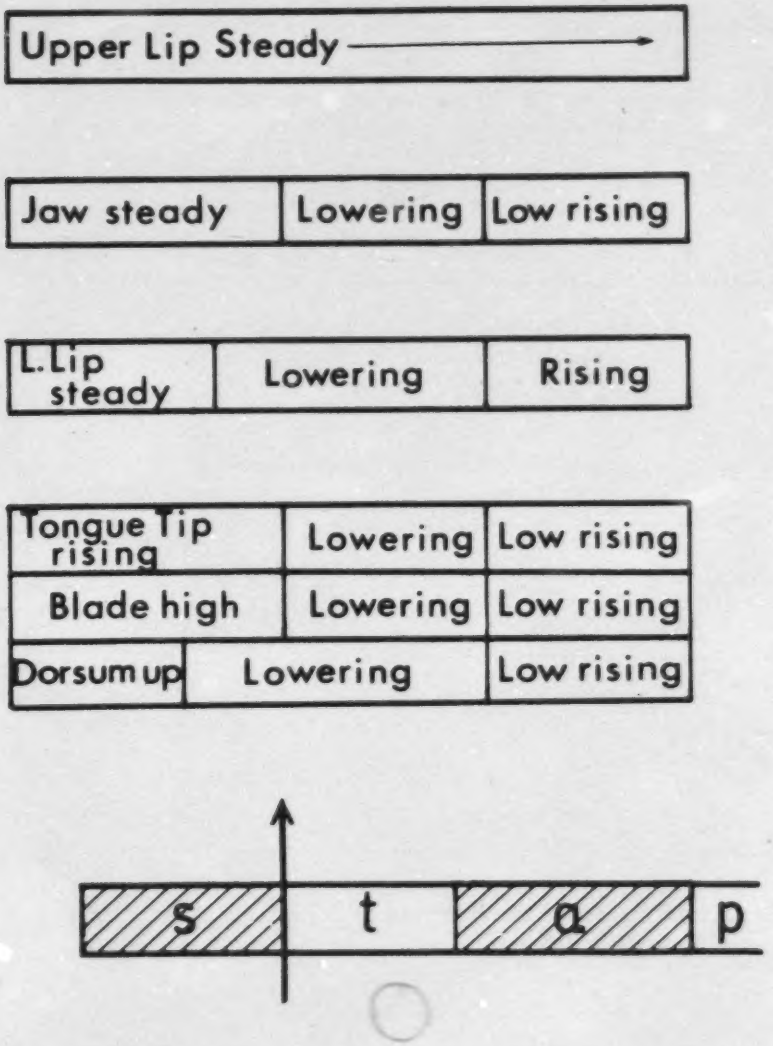


Figure 4: Schematic of differential tongue movement for /stɑpə/. Posterior pellet lowers ahead of anterior pellets.

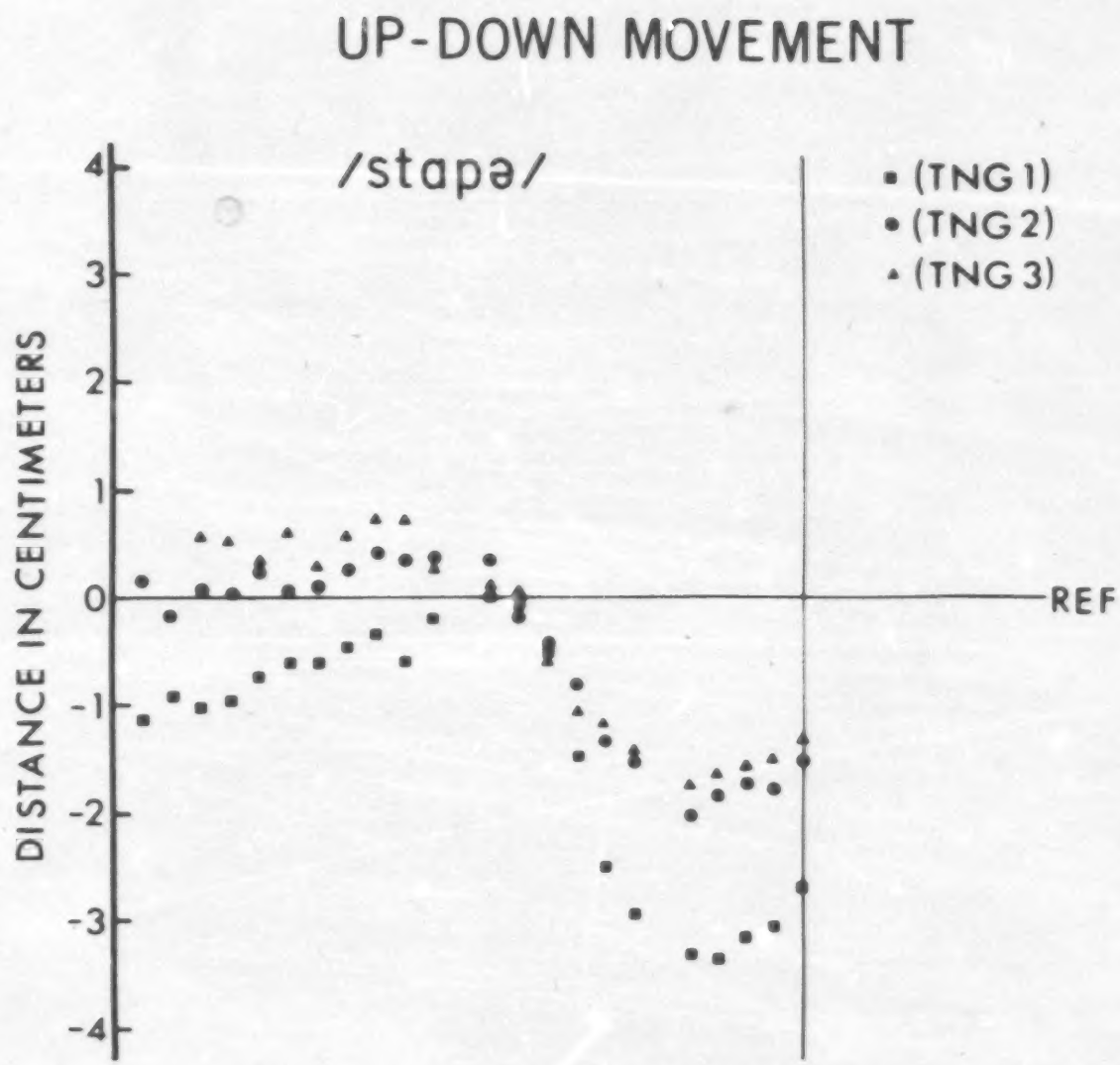


Figure 5: Raw data of pellet positions for one subject during utterance /stapə/.

pellet, however, represented by the small triangle, can be seen to start its descent for /ɑ/ during the /st/ production, in this case three frames or 50 msec before the more anterior portions of the tongue are free to move. The raw data presented in the figures have not been smoothed. The occasional jumps in data points are spurious and not physiological.

Figure 6 shows that for /skɑpə/ as well the tongue moves differentially and not as a unit. In this case the more posterior part of the tongue is involved with the stop and the anterior portion with the fricative, so that while the tongue tip and blade are up for the /s/ constriction, the dorsum is rising for /k/ and is not free to lower until the release of the occlusion. Note, however, that first the tip lowers, then the blade. The tongue-tip-low subject followed the same strategy as the other subjects with the exception that the tip remained low, lowering further simultaneously with the blade as the more posterior dorsum of the tongue concluded the /k/ gesture. The jaw movement is interesting in the case of the /sk/ cluster. Recall that for /sp/ and /st/ the jaw remained high during /s/, whereas here for /sk/ it starts to lower during the fricative. Not only do the high lip position necessary for /p/ and the high tongue position necessary for /t/ put constraints upon jaw lowering, but jaw lowering may facilitate the elevation of the back of the tongue necessary for the /k/. To take one example of the differential tongue movement for /sk/, it can be seen in Figure 7 that the more anteriorally placed pellets lower before the dorsal pellet, in the case of the tip, by 50 msec and the blade, by 17 msec. Figure 8 represents horizontal tongue movement. This is the subject who used more fronting and backing of the tongue than the other subjects. Here it can be seen that the tongue tip started back for /ɑ/ 83 msec before the dorsum and the blade, 33 msec before the dorsal pellet. It seems then that when the tongue is not involved in the stop for an /s/-stop cluster, it is free to move toward the vowel immediately after its involvement with /s/ production in /sp/, whereas for /st/ and /sk/, since the tongue is involved in both parts of the cluster, only the uninvolved portions of the tongue are free to move toward the vowel position during the consonant cluster production.

The strategies used by all three subjects for /st/ and /sk/ indicate a remarkable economy of effort in that only those portions of the tongue primarily involved in a high tongue gesture remain elevated. Other parts of the tongue lower for the following vowel as soon as they are free to do so; for /st/, the more anterior portion of the tongue is involved for both /s/ and /t/, but the dorsum lowers early; whereas for /sk/, the reverse is true, with the anterior portions of the tongue lowering for the vowel during the dorsal elevation for the /k/. These observations are compatible with Öhman's (1966) ideas about consonant and vowel articulation; that is, movements for the vowel can occur during production of the preceding consonant.

It has been reported by Schwartz (1970) and by Klatt (1971), among others, that the duration of the acoustic noise for /s/ is shortened when clustered with a stop, and also that there is a tendency for the /s/ in /sp/ to be shorter than the /s/ in /st/ or /sk/. To assess acoustic durations, spectrograms of all the subjects' speech were measured to the nearest one-tenth of an inch and the measurements were converted into milliseconds. The results are shown in Figure 9. The subjects were consistent in the shortening of the /s/ before the labial stop relative to the palatal stops. The differences were approximately 20 and 10 msec for the first subject, 9 msec for the second, and 18 and 37 msec for the third

/skapə/

Strategy: Tongue starts lowering (and backing) tip first, then anterior blade, then post blade.

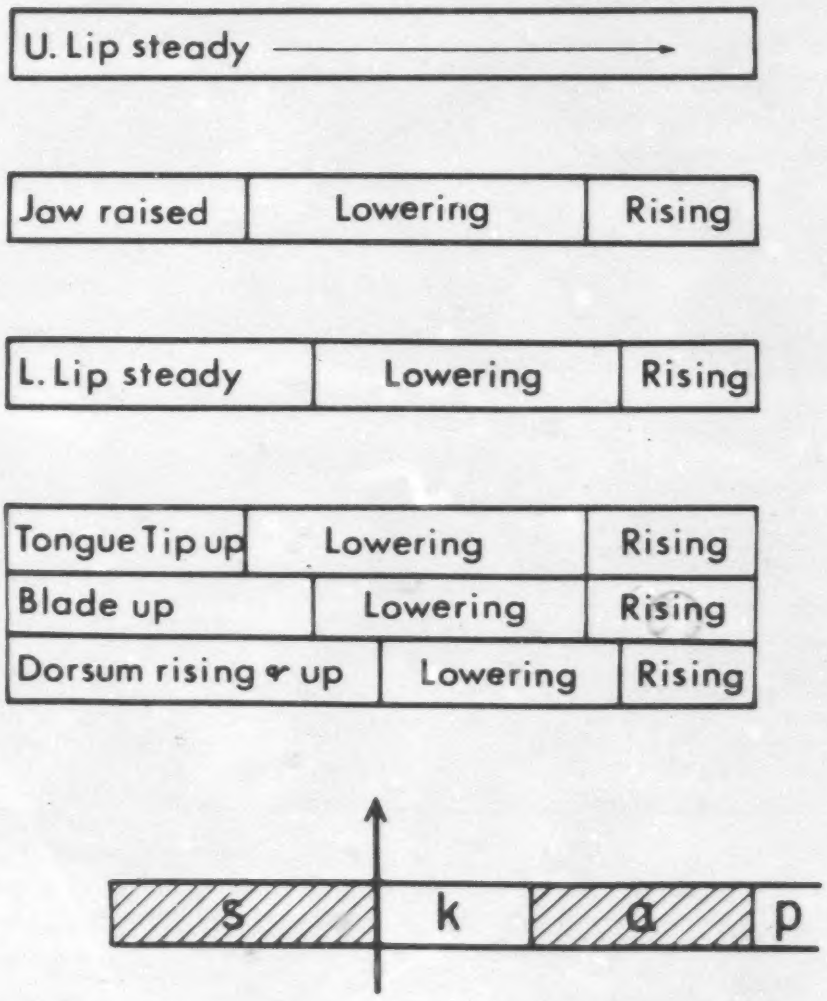


Figure 6: Schematic of differential tongue movement for /skapə/. Anterior pellets lower before posterior pellet.

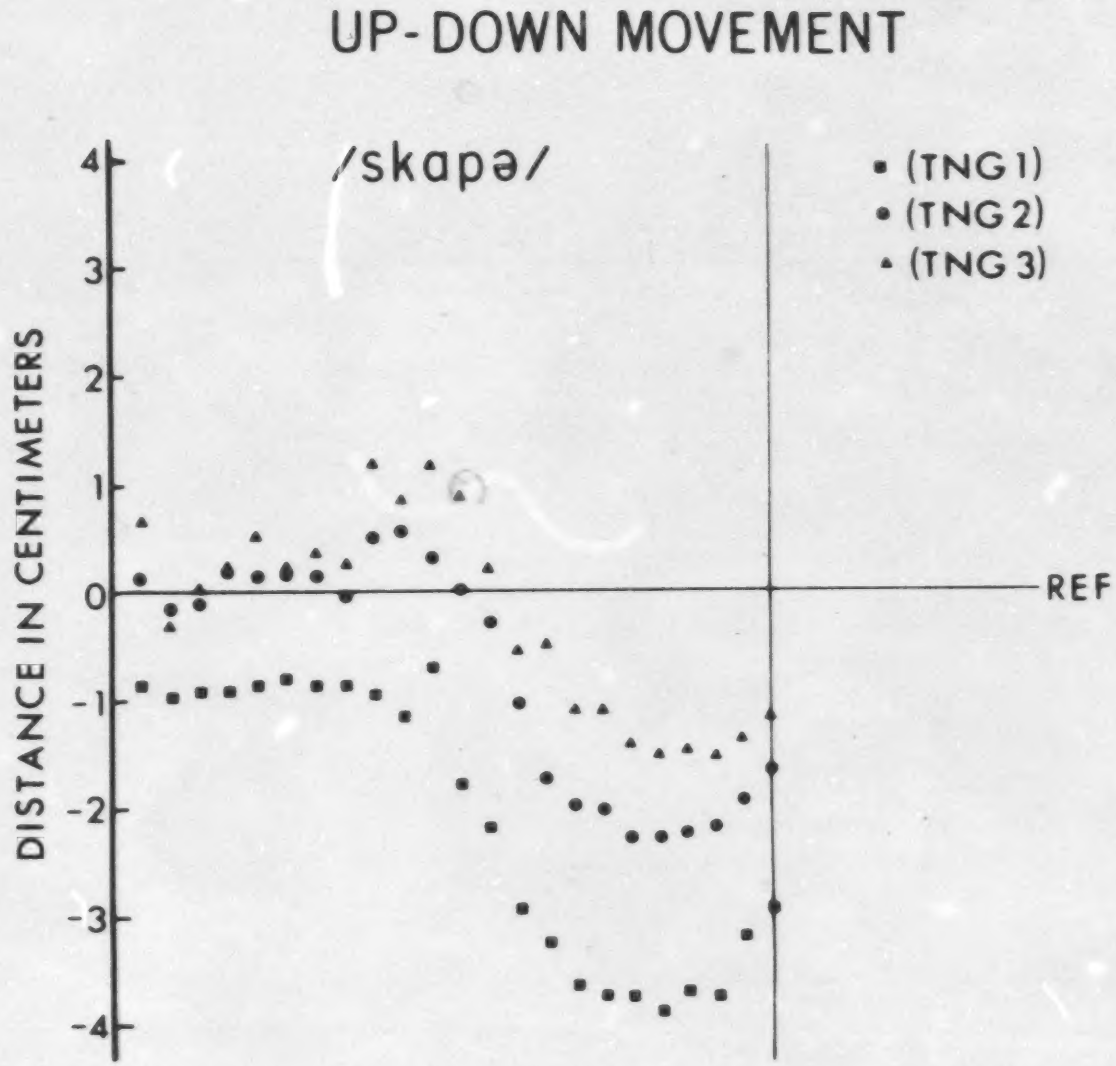


Figure 7: Raw data of pellet positions for one subject during utterance /skɑpə/.

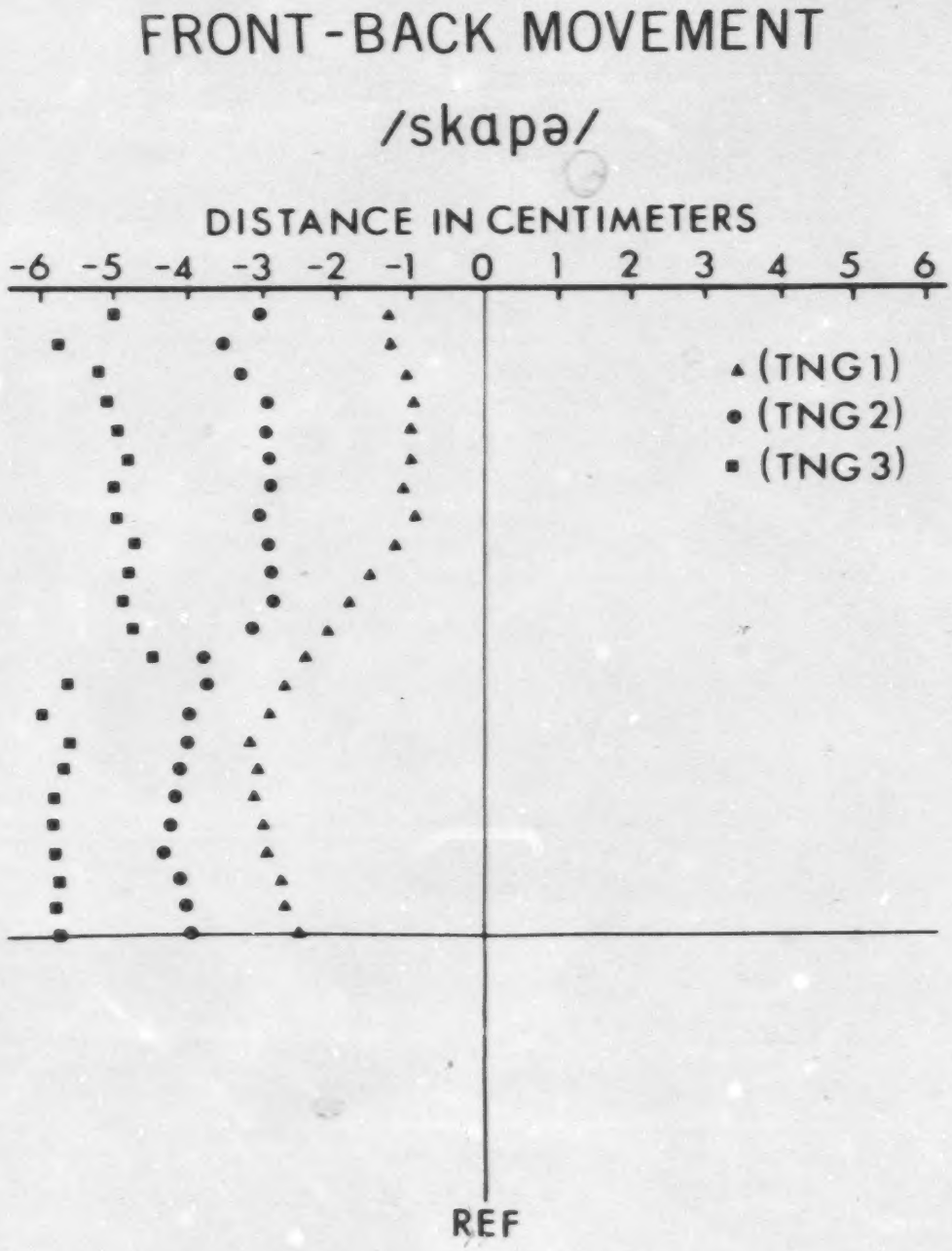


Figure 8: Raw data of horizontal tongue pellet movement for one subject during utterance /skɑpə/.

subject. Note that the third subject, GS, was remarkably consistent in both stop-closure duration and vowel durations.

	s	p	a		s	t	a		s	k	a
GB	130	112	168		150	70	200		140	103	200
FC	94	112	140		103	131	131		103	103	159
GS	94	75	150		112	75	150		131	75	150

Figure 9: Acoustic durations in msec.

Figure 10 relates the acoustic data to the movement data by showing the critical movements for the stop closure for /p/, /t/, and /k/. The subject in this figure is the one who conveniently produced stop closures of 75 msec and vowels of 150 msec across these clusters. Using the closure of the lips after /ɑ/ as the lineup point for the movement comparisons, we can see that the back pellet (marked T. Dorsum) rises during /s/, continues rising to a maximum, during the /k/ closure, and then moves down for /ɑ/. The pellet on the tongue tip seemed critical for the /t/, and the pellet on the lower lip, for the /p/. The difference in timing is evident here but it is easier to compare them if we track from left to right by changing the lineup point.

In Figure 11, we have lined up the tongue- and lip-movement curves at the onset of /s/, since that is the segment differing in duration, and it is obvious that the movement toward closure for the stop is temporally different for /p/, /t/, or /k/. The peak amplitude for the lower-lip movement for /p/ occurs closer in time to the onset of the /s/ noise than does the peak amplitude of movement of the tongue for either /t/ or /k/.

The /s/ seems to be shorter in the case of /sp/, not because the lips close with any more force or velocity, but because they start earlier and are not involved in any conflicting gesture. The lips are free to close. The tongue, however, is involved with the /s/ constriction and movement toward closure is delayed.

This effect is reflected in the EMG data recorded from the first subject. Unfortunately, we are unable to include the EMG data in this paper owing to some inconsistencies that must be rechecked. However, we can report that the orbicularis oris muscle is active in the case of /sp/ before the superior longitudinal or styloglossus muscles show activity for the /t/ and /k/ closures, respectively. The observation is contaminated by the fact that, for that subject at least, the orbicularis oris is active not only for the /p/ but for the /s/.

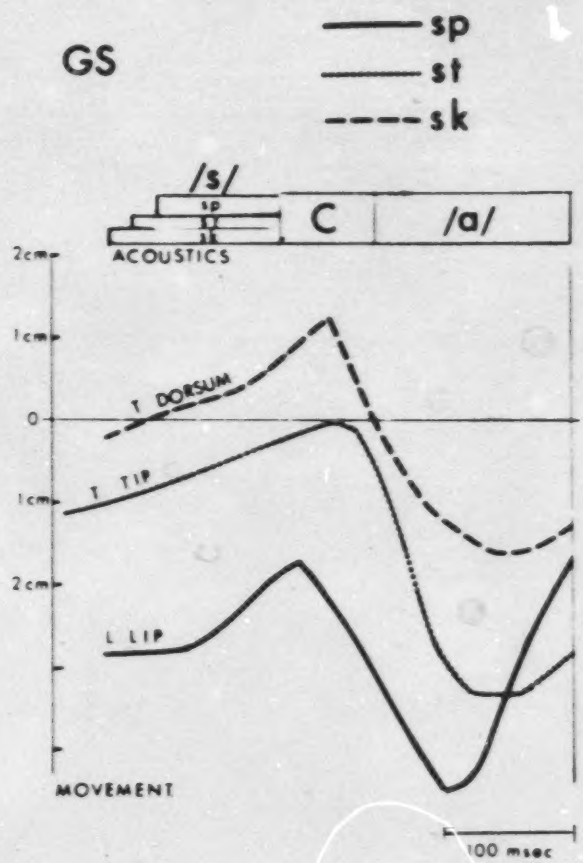


Figure 10: Movements of pellets critical for /p/, /t/, and /k/ closure plotted against acoustic signal.

There are many problems in this combined-techniques approach. One is obtaining clear spectrograms in the presence of the noise of EMG and cine X-ray recording. The measurements are difficult and possible only because of the automatic gain control of the voiceprint that suppresses the background noise in the presence of a strong speech signal.

Another problem in making durational measurements is that both movement and EMG signals are difficult to segment. A tongue blade can move forward and up for both the /s/ and /t/, and one cannot determine clearly the durations of events attributable to each phoneme. For EMG signals there is a problem similar to the one we alluded to in the case of the lip muscles for /s/ and /p/.

Finally, some of the extremely fast adjustments that we find interesting are difficult to capture with techniques that only sample every 20 msec. Despite the difficulties, however, we feel this is an effort worth pursuing in order to explore the temporal aspects of what may be firmly time-locked speech events. We plan another simultaneous EMG and X-ray experiment to further that effort.

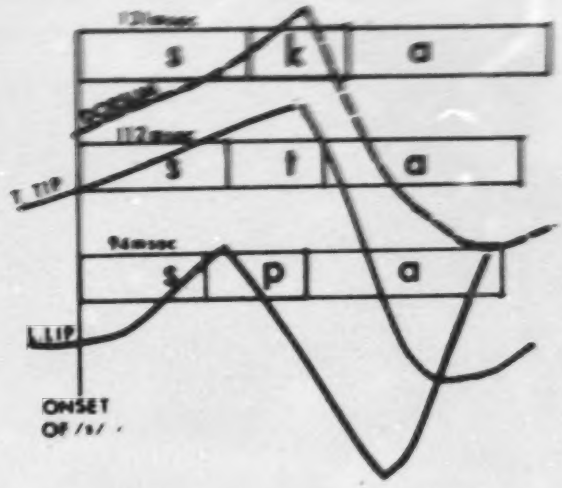


Figure 11: Movements of pellets critical for /p/, /t/, and /k/ closure plotted against acoustic signal. Lip closure for /p/ precedes tongue closure for /t/ or /k/.

REFERENCES

Borden, G. J. and T. Gay. (1975) A combined cinefluorographic-electromyographic study of the tongue during the production of /s/: Preliminary observations. Haskins Laboratories Status Report on Speech Research SR-41, 197-205.

Klatt, D. H. (1971) On predicting the duration of the phonetic segment [s] in English. Quarterly Progress Report (Research Laboratory of Electronics, MIT) QPR-103, 111-126.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Am. 39, 151-168.

Schwartz, M. F. (1970) Duration of /s/ in /s/-plosive blends. J. Acoust. Soc. Am. 47, 1143-1144(Letter).

II. PUBLICATIONS AND REPORTS

III. APPENDIX

Dorman, M. F. and R. J. Porter, Jr. (1975) Hemispheric lateralization for speech perception in stutterers. Cortex 11, 181-185.

Freeman, F. J. (1974) Fluency and phonation. In Vocal Tract Dynamics and Dysfluency, ed. by M. Webster and L. Furst. (New York: Speech and Hearing Institute).

Freeman, F. J., T. Ushijima, and H. Hirose. (1975) Reply to Schwartz's "The core of the stuttering block." Journal of Speech and Hearing Disorders 15, 137-139.

Mermelstein, P. (1975) Book review of G. Fant, Speech Sounds and Features. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-23, 596.

Mermelstein, P. (1975) Vowel perception in consonantal context. Journal of the Acoustical Society of America, Suppl. 58, S56(A).

Nye, P. W., F. S. Cooper, and P. Mermelstein. (1975) Interactive experiments with a Digital Pattern Playback. Journal of the Acoustical Society of America, Suppl. 58, S105(A).

Raphael, L. J., M. F. Dorman, F. J. Freeman, and C. Tobin. (1975) Vowel and nasal duration as cues to voicing in word-final stop consonants: Spectrographic and perceptual studies. Journal of Speech and Hearing Research 18, 389-400.

Repp, B. H. (1976) Dichotic "masking" of voice onset time. Journal of the Acoustical Society of America 59, 183-194.

Turvey, M. T. (1975) Perspectives in vision: Conception or perception? In Reading, Perception and Language, ed. by D. Duane and M. Rawson. (Baltimore, Md.: York).

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers:

SR-21/22 to SR-42/43

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October- December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445
SR-39/40	July - December 1974	AD A007342	ED-102-633
SR-41	January - March 1975	AD A103325	ED-109-722
SR-42/43	April - September 1975	AD A018369	

AD numbers may be ordered from: U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from: ERIC Document Reproduction Service
Computer Microfilm International Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.